University of Alberta

Molecular Characterization of Petroleum Mixtures

by

Zhanyao (Harry) Ha

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Chemical Engineering

Department of Chemical and Materials Engineering

Edmonton, Alberta

Fall 2005

Library and Archives Canada

Published Heritage Branch

395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque et Archives Canada

Direction du Patrimoine de l'édition

395, rue Wellington
Ottawa ON K1A 0N4
Canada

0-494-08650-5

Your file Votre référence
ISBN
Our file Notre reference
ISBN

NOTICE:
The author has granted a non-exclusive license allowing Library and Archives Canada to reproduce, publish, archive, preserve, conserve, communicate to the public by telecommunication or on the Internet, loan, distribute and sell theses worldwide, for commercial or non-commercial purposes, in microform, paper, electronic and/or any other formats.

The author retains copyright ownership and moral rights in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

AVIS:
L'auteur a accordé une licence non exclusive permettant à la Bibliothèque et Archives Canada de reproduire, publier, archiver, sauvegarder, conserver, transmettre au public par télécommunication ou par l'Internet, prêter, distribuer et vendre des thèses partout dans le monde, à des fins commerciales ou autres, sur support microforme, papier, électronique et/ou autres formats.

L'auteur conserve la propriété du droit d'auteur et des droits moraux qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

In compliance with the Canadian Privacy Act some supporting forms may have been removed from this thesis.

While these forms may be included in the document page count, their removal does not represent any loss of content from the thesis.

Conformément à la loi canadienne sur la protection de la vie privée, quelques formulaires secondaires ont été enlevés de cette thèse.

Bien que ces formulaires aient inclus dans la pagination, il n'y aura aucun contenu manquant.

Canada

Abstract

Petroleum mixtures contain numerous types of hydrocarbon molecules. Characterizations of such complex mixtures are critical to kinetic modeling during refining, but providing such characterizations is a challenge. The isomeric lump in homologous series sets up the analytical limit for characterizing the composition of heavy petroleum oil. By minimizing the Gibbs free energy of the whole isomeric lump, subject to the stoichiometric constraint and the measured averaged boiling point of that lump, an optimization approach has been proposed to quantify the compositional distribution among the isomers. This thesis presents a computer-assisted deterministic approach for automatically generating hydrocarbons, optimizing the molecular geometry on-line, selecting the most possible molecules, distributing them within a measured isomeric lump, predicting the physical and thermodynamic properties of each molecule using Quantitative-Structure-Property-Relationship (QSPR) models, and, consequently, processing them to predict the properties of a mixture. The simulated results were compared with five diesel samples of measured properties for molecular weight, density, refractive index, and simulated distillation curves. Good agreements were found between predicted bulk properties and measured ones for all five diesel samples, and thus, these results indicated the proposed methodology could be used to derive the detailed molecular representation for middle distillates.

A new approach has been proposed to quantify the compositional distribution of different hydrocarbon isomers in an "isomeric lump" determined using gas

chromatography – mass spectrometry (GC-MS) methods. It appears that by minimizing the Gibbs free energy of the mixture of the major isomers, subject to the stoichiometric constraint and the measured average boiling point of that isomeric lump, the concentration distribution of various isomers can be determined with good accuracy. The simulated compositions of the heptane isomers were compared with the reported analytical results for 18 crude oils. The correspondence between predicted and measured distributions was found to be good. The reported distributions of the heptane isomers are far from the thermodynamic equilibrium distribution, so the introduction of an average boiling point constraint helps these distributions match. With this delumping approach, the petroleum mixtures can be characterized at a molecular level, which is beyond the analytical limitations.

Acknowledgements

I would like to thank my supervisors Dr Shijie Liu and Dr Zbigniew Ring for their support and guidance throughout my doctoral program. Their knowledge, enthusiasm, and attention to details are reflected throughout this thesis. I wish to thank my committee members, Dr Zhenghe Xu, Dr R. E. Hayes, and Dr Liang Li, for their encouragement and suggestions.

Thank all the members in the Secondary Upgrading & Refining Group of National Center for Upgrading Technology (NCUT), Natural Resources Canada, for providing an enjoyable environment for research and learning. In particular, Norma, Cecilia, Jinwen, Hong, Renata, Jenny, and Govind had been a great help.

This work was conducted under the funding of NCUT provided by the Canadian Program for Energy and Development, the Alberta Research Council and the Alberta Energy Research Institute.

My deepest thanks to my wife, Huan, for her love, encouragement, and sacrifice throughout my Ph.D study.

Table of Contents

Chapter 1. Introduction 1

1.1 Background 1

1.2 Research Objectives 6

1.3 Thesis Outline 8

1.4 References 9

Chapter 2. QSPR Models for Boiling Points, Specific Gravities and Refraction Indices of Hydrocarbons 12

2.1 Introduction 12

2.2 Data sets 15

2.3 Methodology 17

2.3.1 Molecular descriptors 17

2.3.2 Molecular modeling and input file preparation 18

2.3.3 Descriptor screening and regression analysis 18

2.3.4 Model validation 20

2.3.5 Inter-correlation of the modeled properties 21

2.4 Results and Discussion 22

2.4.1 Modeling of BP for saturates 22

2.4.2 Modeling of BP for aromatics 26

2.4.3 Modeling of BP for both saturates and aromatics 28

2.4.4 Modeling of SG for saturates 30

2.4.5 Modeling of SG for aromatics 32

2.4.6 Modeling of SG for both saturates and aromatics 35

2.4.7 Modeling of RI for saturates 36

2.4.8 Modeling of RI for aromatics 39

2.4.9 Modeling of RI for both saturates and aromatics 41

2.5 Conclusions 43

2.6 References 44

Chapter 3. Hydrocarbon Isomer Distribution in Petroleum Mixtures 49

3.1 Introduction 49

3.2 Isomer distribution within an isomeric lump 52

3.3 Simulated isomer distributions and discussion 55

3.4 Conclusions 64

3.5 References 65

Chapter 4. Data Analysis and Reconciliation 68

4.1 Introduction 68

4.2 Compositional analyses and bulk property measurements 69

4.3 Integration of PIONA and GC-FIMS results 72

4.4 Data reconciliation between GC-FIMS and SimDis 76

4.5 Conclusions 82

4.6 References 82

Chapter 5. Derivation of Molecular Representations of Diesels 84

5.1 Introduction 84

5.2 Methodology 87

5.2.1 Analytical measurements and data reconciliation 89

5.2.2 Molecular generation and simulation 89

5.2.2.1 Molecular generation rules 89

5.2.2.2 Molecular fingerprint in MOPAC input format 95

5.2.2.3 Molecular simulation, selection, thermodynamic property 97

5.2.2.4 Generation and manipulations of hydrocarbon molecules 99

5.2.2.4.1 Molecular generation of paraffins 102

5.2.2.4.2 Molecular generation of cyclic hydrocarbons 105

5.2.3 QSPR study of BP, SG, an RI 111

5.2.4 Isomer distribution within an isomeric lump 112

5.2.5 Derivation of minimum set molecular representation 114

5.3 Molecular representation and bulk property predictions 116

5.4 Sensitivity of properties to the size of molecular representation 120

5.5 Conclusions 121

5.6 References 123

Chapter 6. Conclusions and Recommendations 128

6.1 Conclusions 128

6.2 Recommendations 131

Appendix A. The raw data set for QSPR correlations 132

Appendix B. The derivation of uniqueness for quasi-equilibrium isomer distribution

143

Appendix C. Figures C1-C14. Comparisons of predicted distributions of heptane isomers with the reported data in 18 crude oils 146

Appendix D. PIONA, GC-FIMS, and data reconciliation results for samples 2 through 5 154

Appendix E. A comprehensive comparison between this work and reported characterization methods 162

List of Tables

Table 2.1 Correlation coefficients between *BP*, *SG*, and *RI* 22

Table 2.2 *BP* model for saturates 24

Table 2.3 *BP* model for aromatics 26

Table 2.4 *BP* model for both saturates and aromatics 29

Table 2.5 *SG* model for saturates 30

Table 2.6 *SG* model for aromatics 33

Table 2.7 *SG* for both saturates and aromatics 35

Table 2.8 *RI* model for saturates 37

Table 2.9 *RI* model for aromatics 39

Table 2.10 *RI* model for both saturates and aromatics 42

Table 3.1 Abundances of heptane isomers in virgin crude oils (Tissot and Welte, 1984) 53

Table 3.2 Properties of hexane/heptane isomers and average predicted results 56

Table 3.3 General description of 18 crude oils (Martin et al., 1963) 57

Table 3.4 The prediction deviations (calculated-measured) for normalized distribution of hexane/heptane isomers (wt%) 59

Table 4.1 Summary of the bulk properties and main components for five diesel samples 70

Table 4.2 PIONA report for sample 1 71

Table 4.3 GC-FIMS by #C distribution report normalized for >200°C fractions for sample 1 72

Table 4.4 Integrated results from PIONA and GC-FIMS measurements for sample 1 72

Table 4.5 Paraffinic calibration (RT-BP) check for SimDis and GC-FIMS 78

Table 5.1 Branch types being attached on main chain or main ring core 102

Table 5.2 QSPR models' performance on BP, SG, and RI predictions for pure hydrocarbons 112

Table 5.3 Simulated bulk properties compared with experimental results for five samples 117

Table 5.4 Simulated results with minimum set of representative molecules 121

List of Figures

Figure 1.1 The evolution of refinery technology 1

Figure 2.1 Parity plot of the BP model for saturates; 186-point training set 24

Figure 2.2 Comparison of the BP model for saturates with Joback's group contribution model, Parity plot for the 34-point test set 25

Figure 2.3 Parity plot of the BP model for aromatics; 200-point training set 27

Figure 2.4 Comparison of the BP model for aromatics with Joback's group contribution model, Parity plot for the 61-point test set 28

Figure 2.5 Parity plot of the BP model for saturates and aromatics; 386-point training set 29

Figure 2.6 Parity plot of the SG model for saturates; 186-point training set 31

Figure 2.7 Comparison of the SG model for saturates with Rackett's group contribution model, Parity plot for the 34-point test set 32

Figure 2.8 Parity plot of the SG model for aromatics; 200-point training set 34

Figure 2.9 Comparison of the SG model for aromatics with Rackett's group contribution model, Parity plot for the 36-point test set 34

Figure 2.10 Parity plot of the single SG model for saturates and aromatics; 386-point training set 36

Figure 2.11 Parity plot of the *RI* model for saturates; 186-point training set 38

Figure 2.12 Parity plot of the *RI* model for saturates with the 34-point test set 38

Figure 2.13 Parity plot of the *RI* model for aromatics; 200-point training set 40

Figure 2.14 Parity plot of the *RI* model for aromatics with the 27-point test set41

Figure 2.15 Parity plot of the single *RI* model for saturates and aromatics; 386-point training set 42

Figure 3.1 Prediction of heptane isomer distribution in Alida crude oil 60

Figure 3.2 Prediction of heptane isomer distribution in South Houston crude oil 61

Figure 3.3 Prediction of heptane isomer distribution in Wafra crude oil 62

Figure 3.4 Prediction of heptane isomer distribution in Wilmington crude oil 62

Figure 4.1 A programming flow chart for PIONA-FIMS integration 73

Figure 4.2 A programming flow chart for data reconciliation between GC-FIMS and SimDis 77

Figure 4.3 Data reconciliation result for sample 1 79

Figure 4.4 Data reconciliation result for sample 2 80

Figure 4.5 Data reconciliation result for sample 3 80

Figure 4.6 Data reconciliation result for sample 4 81

Figure 4.7 Data reconciliation result for sample 5 81

Figure 5.1 Molecular delumping algorithm for molecular characterization 88

Figure 5.2 Thermodynamic stabilities of C9 isoparaffins 91

Figure 5.3 Molecular simulations of C10 alkyl-benzenes 95

Figure 5.4 The internal coordinates (MOPAC input format) of ethane 97

Figure 5.5 Programming diagram for molecular generation 100

Figure 5.6 Programming diagram for isoparaffin generation 104

Figure 5.7 Ring-core structures of cyclic hydrocarbons representing molecules identified by GC-FIMS 106

Figure 5.8 Programming diagram for generation of cyclic hydrocarbons 109

Figure 5.9 GC-FIMS-generated SimDis and simulated SimDis compared with measured SimDis for sample 1 118

Figure 5.10 GC-FIMS-generated SimDis and simulated SimDis compared with measured SimDis for sample 2 118

Figure 5.11 GC-FIMS-generated SimDis and simulated SimDis compared with measured SimDis for sample 3 119

Figure 5.12 GC-FIMS-generated SimDis and simulated SimDis compared with measured SimDis for sample 4 119

Figure 5.13 GC-FIMS-generated SimDis and simulated SimDis compared with measured SimDis for sample 5 120

Chapter 1

Introduction

1.1 Background

Over the last century, the refining industry experienced three evolutionary periods: the Separations and Thermal period, the Catalytic period, and the Quantitative Reaction Engineering period as shown in Figure 1.1 (Katzer et al., 2000).

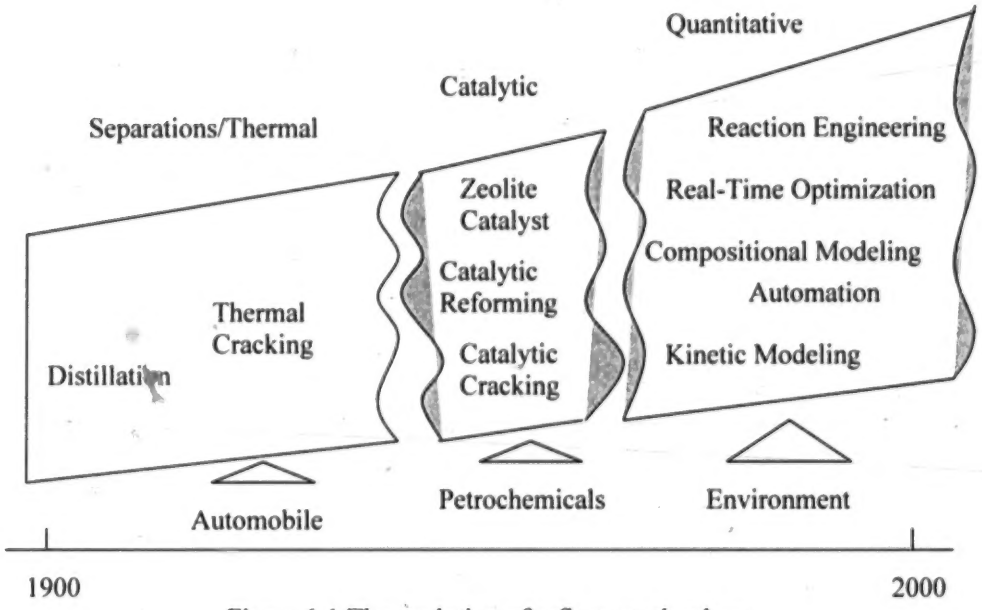


Figure 1.1 The evolution of refinery technology

Three factors have played important roles in this evolution. The first is that crude oils have become heavier and heavier and the sulfur contents are getting higher and higher. The second is that the demand for heavy fuels, as used in power stations and industrial burners, has substantially decreased because of global energy conservation efforts and a switch to alternative energy sources. As a result, the demand of secondary upgrading processes such as FCC (Fluid Catalytic Crack) and hydrocracking has increased. The

third factor is that environmental concerns are calling for cleaner oil products, and the product specifications are becoming stricter. For example, Europe, the United States, and Japan will lower the allowable sulfur content in commercial fuels to <20 *ppmw* in the near future. Therefore, refineries are now facing big challenges: eating worse but producing better. One recipe for this problem is to have a better understanding for the chemistry and chemical reactions involved for the feedstock, effluent, as well as the streams in processing. This calls for detailed molecular mechanistic modeling, instead of traditional lumped modeling, for the kinetic studies of refinery processes.

As the key starting point, the characterization of petroleum fractions is essential in refinery modeling. The lack of an adequate heavy oil characterization had been considered the bottleneck for refinery process integration (Vandereijk et al., 1990). Developments of new process models rely heavily on advances in characterization methods, and developing characterization methods, in turn, relies on the advance of analytical techniques. The evolution of FCC modeling over the past 40 years illustrates these dependencies. Petroleum mixtures can consist of order of 10^5 molecular compounds; measuring and characterizing each molecule is infeasible. Therefore, some level and type of lumping are inevitable. A traditional method is to lump the huge number of molecules into manageable lumps. These lumps are taken as reactants whose physical and chemical properties are identical to that of the weighted sum of their individually contained components. With limited analytical technology, a three-lump model was developed by Weekman and Nace (1970). Later, more chemistry information was incorporated as the number of lumps increased to 4 (Oliveira and Biscaia, 1989), 5

(Ancheyta-Juarez et al., 1999), and 10 (Jacob et al., 1976). Lumped models are fast and easy to solve. They are still dominant in process control, design, and optimization.

However, when lumping molecules, the chemical mechanisms of individual molecules are obscured and cannot be incorporated into the kinetic models. A lumped model is usually compositional dependent and fails in prediction when the feedstock changes. Crude oil composition can affect product yield and quality, and market prices of products can influence operating strategies. A refinery process might therefore require frequent, compositionally dependent optimizations. Further, new environmental air-quality regulations place restrictions on the molecular composition of petroleum products, in addition to their traditional physical properties. Therefore, there is a need for improved understanding of complex reaction mixtures, and for rigorous process models that are independent of the feedstocks and more molecularly explicit.

With the advances in analytical chemistry and computer technology, a complex mixture can now be modeled at the level of isomeric lumps. The Structural Oriented Lumping (SOL) model of Mobil (Quann and Jaffe, 1992, 1996) lumped the hydrocarbons into 95 molecular classes, which in turn are expanded into more than 3000 hydrocarbon molecules. A strong assumption when lumping a petroleum mixture into homologous series is that the physical and chemical properties of the hydrocarbon isomers are identical. Reaction rules (e.g. ring opening, sulfur removal) derived from fundamental reaction chemistry are then applied to track the changes of the set of structural vectors, and thus build the reaction network. The SOL approach opened a door to develop the molecular-based models of an entire refinery since all the petroleum processes have the same representation of a complex mixture. The approach was applied to model catalytic

hydroprocessing (Quann and Jaffe, 1992) and catalytic cracking (Christensen et al., 1999) processes. However, the details of this approach are not available in the open literature.

Froment and co-coworkers developed detailed kinetic models that track surface intermediates and fundamental molecular transformations for hydrocracking (Baltanas and Froment, 1985), thermal cracking (Clymans and Froment, 1984), as well as catalytic cracking (Dewachtere et al., 1999). Using the Boolean matrices as logical operators to describe the structure of hydrocarbon molecules and intermediates, and to account for each step of the reaction mechanism, a complete reaction network was generated automatically. The rate of each elementary reaction was then determined using kinetic correlations based on Single-Event theory (Feng et al., 1993). Basically, the fundamental mechanistic model requires the precise determination of the mixture composition. However, no molecular characterization method was employed to facilitate their model. Instead, the measured isomeric lumps (by #C) were used to build up the reaction network.

Klein and co-workers carried out research on mechanistic modeling for the reaction of catalytic cracking (Joshi et al., 1997), hydroprocessing (Hou and Klein, 1999), and pyrolysis (Fake et al., 1997). These models are based on the fundamental reaction steps such as β-scission, hydride shift, methyl shift, bond fission etc., which involve elementary transitions of active centers. Based on the Evans-Polanyi structure-reactivity theory (Evans and Polanyi, 1938), the kinetic parameters are correlated with the "reaction index" using Linear Free Energy Relationship correlations (Watson et al., 1996). The "reaction index" can be a property of the reactants or a property of the reaction. Molecules generated by Monte Carlo Simulation were used to characterize petroleum mixtures (Neurock et al., 1994), and results were incorporated in the authors' model

development. Probability distribution functions were derived from the bulk properties to represent various structural attributes of petroleum mixtures. These attributes were then sampled randomly to build a set of molecules to fit the measured bulk data. Due to the randomness, a large set of $O(10^5)$ molecules had to be generated for better accuracy. Efforts were made to reduce the number of molecules and to meet the needs of kinetic modeling (Campbell and Klein, 1997).

However, completely random molecular-generation can easily introduce redundant molecules, and the molecules generated cannot be guaranteed to exist in real petroleum mixtures. Another problem associated with this stochastic method is that the probability functions derived from the average structural parameters may not provide the actual compositions in petroleum mixtures. Kim et al. (1998) investigated the molecular compositions of three FCC feedstocks that have very similar bulk properties. The Mass Spectrometry (MS) analysis revealed that the molecular type distributions were significantly different among the three feeds. Ramaswamy et al. (1989) also found that the hydrocarbon type distributions were entirely different between two VGOs although the two samples had almost the same SARA (saturates, aromatics, polars, and asphaltenes) analysis results. Therefore, molecules completely constructed from the average structural parameters may lead to an uncertainty in the composition of petroleum mixtures. Different molecular structures of FCC feedstocks have different cracking patterns; therefore, the product yields and properties might be totally different, even though such feedstocks have very similar bulk properties. In the cases when the detailed compositional analytical data (e.g. GC-FIMS or LC-FIMS) are available, more accurate and efficient characterization methods should be used.

The mechanistic models provide the most direct expression of underlying physical and organic chemistry and, hence, the kinetic parameters are independent of feedstocks and the processing conditions. They also have molecular outputs, which allow the predictions of product properties. These models can meet the future environmental and industrial requirements in process design, quality control, and process optimization. Applying mechanistic models needs precise molecular characterization methods, which are not available today. The ultimate goal of this work is to develop a molecular characterization with an adjustable number of molecules that can describe the petroleum mixtures with high accuracy and detail.

1.2 Research Objectives

The goal of this project was to develop a consistent molecular characterization for petroleum mixtures to facilitate the development of modeling both reaction (e.g., hydrotreating, hydrocracking, catalytic cracking etc.) and separation processes. The proposed characterization method is more molecularly explicit and flexible than the existing ones. It can also be used to predict the bulk properties of petroleum mixtures when applying appropriate mixing rules. To assist the development and check the validity of the characterization method, the following investigations are requisite:

1). To develop accurate Quantitative-Structure-Property-Relationship (QSPR) models for the estimations of Boiling Point (BP), Specific Gravity (SG), and Refractive Index of pure hydrocarbon compounds. These QSPR models are used to predict the above physical properties for each individual molecule generated. As a result, the bulk

properties such as density, RI, and Simulated Distillation (SimDis) can be predicted from the pure components once the molecular representation of a mixture is known.

2). To develop a distribution method for hydrocarbon isomer distributions. This thesis uses the GC-FIMS report (hydrocarbon type distributed by carbon number) as the starting point for molecular characterization. Each cell in a GC-FIMS report represents an isomeric lump corresponding to a hydrocarbon type and a carbon number. This isomeric lump has to be delumped to have a molecular representation. Based on the constrained thermodynamic equilibrium theory, a delumping approach is proposed, and thereby, the molecular representation is obtained.

3). To develop a data reconciliation method that will check the consistency between GC-FIMS and SimDis measurements. For some samples containing substantial amounts of light fractions, PIONA analyses are required. Therefore, an integration of PIONA and GC-FIMS results is needed to have a full range distribution of hydrocarbons. The reconciliation procedures are developed as a quality control approach. As a result, consistent and reliable GC-FIMS results can be produced.

4). To design a molecular generation method that will facilitate the molecular simulation on-line. A MOPAC input format is used for molecular generation and MOPAC2002 is applied for on-line molecular geometry optimization and thermodynamic calculations. Procedures regarding the molecular generation, simulation, selection, and distribution will be developed accordingly to complete an automation process for molecular characterization.

Due to the availability of data from GC-FIMS at NCUT (limited to the diesel range), the current work focuses on developing a molecular characterization method to represent

the middle distillates. However, the methodology can be applicable to heavier materials like gas oils. Once the GC-FIMS data is available for heavy materials, it can be directly applied.

1.3 Thesis Outline

Chapter 1 serves as an introduction to the subject of molecular characterization of petroleum mixtures. It is intended to provide the background and motivation for this research and set up the final objectives of this work.

The body of the thesis is in paper-based format from Chapters 2 through 5. Chapter 2 introduces the development of QSPR models for estimating the physical properties of pure hydrocarbons. With the aid of commercial software – CODESSA – separated QSPR models are developed for predicting the BP, SG, and RI for pure saturates and aromatics. The database, methodology, and selection of models and descriptors are described. The advantages of QSPR models for both their accuracy and predicative ability are emphasized as compared with Group Contribution methods.

An isomer distribution approach is proposed in Chapter 3. The thermodynamics governing the distribution of isomers is addressed. The constraints that reflect the distribution of isomers are discussed. The approach is validated by the consistency between the predicted and the reported distributions of heptane isomers. The supporting theory for the proposed approach is also discussed.

Chapter 4 presents a data reconciliation method to reconcile the GC-FIMS results with the SimDis and PIONA data. Integration of PIONA and GC-FIMS measurements is illustrated. The consistency between GC-FIMS and SimDis results is checked for selected

diesels to ensure the data quality for characterization. Procedures that facilitate these QC checks are described in detail.

Chapter 5 is a general application of the above approaches (Chapter 2 to 5), integrated with the molecular generation, simulation, and selection. The choice of using MOPAC input format as the fingerprint of molecules is stated and advantages are justified. Detailed molecular generation procedures are described as incorporated with the on-line simulation, selection, and QSPR predictions. The overall characterization method is applied to 5 diesel samples. The results are presented and discussed.

Chapter 6 is a summary of Chapters 2-5 and includes a discussion relevant to proposed applications of the molecular characterization method. Recommendations for extending current work to the heavier materials like gas oils are made.

1.4 References

Ancheyta-Juárez, J.; López-Isunza, F.; Aguilar-Rodríguez , E. (1999). 5-lump kinetic model for gas oil catalytic cracking, *Applied Catalysis*, **177**, 227-235.

Baltanas, M. A.; and Froment, G. F. (1985). Computer generation of reaction networks and calculation of product distribution in the hydroisomerisation and hydrocracking of paraffins on Pt-containing bifunctional catalyst, *Comput. Chem. Eng.*, **9**, 71-87.

Campbell, D; Klein, M. T. (1997). Construction of a molecular representation of a complex feedstocks by Monte Carlo and quadrature methods. *Applied Catalysis: A General*, **160**, 41-54.

Christensen, G.; Apelian, M. R.; Hickey, K. J.; and Jaffe, S. B. (1999). Future directions in modeling the FCC process: An emphasis on product quality, *Chem. Eng. Sci.*, **54**, 2753-2764.

Clymans, P. J.; and Froment, G. F. (1984). Computer generation of reaction paths and rate equations in the thermal cracking of normal and branched paraffins, *Comput. Chem. Eng.*, **8**, 137-142.

Dewachtere, N. V.; Santaella, F.; and Froment, G. F. (1999). Application of a single-event kinetic model in the simulation of an industrial riser reactor for the catalytic cracking of vacuum gas oil. *Chem. Eng. Sci.*, **54**, 3653-3660.

Evans, M. G.; and Polanyi M. (1938). Inertia and driving force of chemical reactions, *Trans. Faraday Soc.*, **34**, 11-29.

Fake, D. M; Nigam, A.; and Klein, M. T. (1997). Mechanism based lumping of pyrolysis reactions: Lumping by reactive intermediates, *Appl. Cat. A: General*, **160**, 191-221.

Feng, W.; Vynckier, E.; and Froment, G. F. (1993). Single-event kinetics of catalytic cracking, *Ind. Eng. Chem. Res.*, **32**, 2997-3005.

Hou, G.; and Klein, M. T. Automated molecular-based kinetic modeling of complex process-a hydroprocessing application. AIChE National Meeting, Houston, March, 1999.

Jacob, S; Gross, M., B.; Voltz, S. E.; and Weekman, V. W. (1976). A lumping and reaction scheme for catalytic cracking, *AIChE J.*, **22**, 701-713.

Joshi, P.V.; Iyer, S. D.; and Klein, M. T. Computer assisted modeling of gas oil fluid catalytic cracking, 214th National Meeting, ACS, Las Vegas, NV, Sept., 1997.

Katzer, J. R.; Ramage, M. P.; and Sapre A. V. (2000). Petroleum refining: poised for profound changes, *Chem. Eng. Progress*, July, 41-51.

Kim, H. N.; Verstraete, J. P.; Virk, P. S.; and Fafet, A. (1998). NMR enhances mass-spec FCC feedstocks characterization. *Oil & Gas Journal*, **96**, 85-88.

Neurock, M. N.; Nigam, A.; Trauth, D.; and Klein, M. T. (1994). Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms, *Chem. Eng. Sci.*, **49**, 4153-4177.

Oliveira, L. L.; and Biscaia Jr., E. C. (1989). Catalytic cracking kinetic models. Parameter estimation and model evaluation, *Ind. Eng. Chem. Res.*, **28**, 264-271.

Quann, R. J.; and Jaffe, S. B. (1992). Structure-oriented lumping: Describing the chemistry of complex hydrocarbon mixtures, *Ind. Eng. Chem. Res.*, **31**, 2483-2497.

Quann, R. J.; and Jaffe, S. B. (1996). Building useful models of complex reaction systems in petroleum refining, *Chem. Eng. Sci.*, **51**, 1615-1635.

Ramaswamy, V.; Singh, I. D.; Krishna, R. (1989). Characterization of vacuum gas oil from North Gujarat crude mix, *Indian Journal of Technology*, **27**, 85-88.

Vandereijk, H.; Denotter, G. J.; Blauwhoff, P. M. M.; Maxwell, I. E. (1990), The application of advanced process models in oil refining R&D, *Chem. Eng. Sci.*, **45**, 2117-2124.

Watson, B. A.; Klein, M. T.; and Harding, R. H. (1996). Mechanistic modeling of n-Heptane cracking on HZSM-5, *Ind. Eng. Chem. Res.*, **35**, 1506-1516.

Weekman, V. W.; and Nace, D. M. (1970). Kinetics of catalytic cracking selectivity in fixed, moving, and fluid-bed reactors, *AIChE J.*, **16**, 397-404.

Chapter 2

QSPR Models for Boiling Points, Specific Gravities and Refraction Indices of Hydrocarbons

2.1. Introduction

The normal boiling point (*BP*), specific gravity (*SG*), and refractive index (*RI*) are some of the most important physical properties of hydrocarbon compounds. They are good indicators of crude oil quality and can be easily and precisely measured. Many refinery units are controlled using distillation (*BP* distribution) data. Specifications for marketable petroleum products include *BP* distribution and *SG*. Safety and environmental protocols are often associated with these properties. Other physical and chemical properties of hydrocarbon materials are closely related to these three principal properties (Riazi and Roomi, 2001). For example, the n-d-M method (Van Nes and Van Westen, 1951) calculates the aromatic, naphthenic, and paraffinic contents from measurements of *RI*, *SG*, and molecular weight (*MW*). Process models such as the commercial software HYSYS (Aspen Technology, Inc., Calgary) use the above properties as modeling parameters for the estimation of phase equilibrium. Experimental data of *BP*, *SG*, and *RI* are not available for most of the hydrocarbon compounds, especially for larger molecules. Therefore, determination of physical properties of chemical compounds is part of chemists' or engineers' daily routines. Models for accurate prediction of these properties are thus highly desired. This study is aimed at the development of accurate general models for *BP*, *SG* and *RI* of hydrocarbons based on as large a database as possible. The estimation of *RI* has not been studied for pure hydrocarbon compounds, while the available *BP* and *SG* correlations have been developed with limited databases.

It has been generally accepted that the physical properties of a substance are determined by its molecular structure. Various rules and formulae have been proposed to estimate the physical properties of pure hydrocarbon compounds. Of those models, the group contribution methods have been widely used to predict *BP* (Joback, 1984) and density (Elbro et al., 1991). However, Joback's method only works well for hydrocarbons in the intermediate boiling range (300-500K). Poor predictions have been reported beyond this range (Stein and Brown, 1994). The group contribution method for density proposed by Elbro et al. (1991) is accurate only for acyclic alkanes. Moreover, these group contribution methods cannot differentiate between isomers that have exactly the same structural groups but different structural arrangements.

The Quantitative Structure-Property Relationship (QSPR) approach appears to avoid shortcomings of the group contributions methods. It can account for the molecular structure in a more effective way using a variety of descriptors that capture differences between various isomeric structures. In the commercial programs CODESSA (Katritzky et al., 1997) or ADAPT (Stuper et al., 1979), hundreds of descriptors can be generated to represent the molecule. These descriptors are defined based solely on the molecular structure and, in general, represent six different aspects of the molecule: constitution, topology, geometry, electrostatics, quantum-chemistry, and thermodynamics. A QSPR model can be obtained by fitting the available experimental data with a variety of models using those descriptors as variables. In the absence of scientific insight into the causal relationship between the descriptors and the estimated properties, multiple linear regression models or neural networks are used to develop QSPRs and the development becomes an exercise in statistics. QSPR models have been successfully used in such areas

as analytical chemistry and pharmaceutical research (Katritzky et al., 2001; Murugan et al., 1994; Stanton and Jurs, 1990). Models derived solely from the topological index work well for homologous and co-generic series of compounds where the intramolecular interactions parallel the increase in molecular size. Additional molecular descriptors are needed when dealing with a large and diverse set of data (Katritzky and Gordeeva, 1993). Working on a set of 356 hydrocarbon compounds with ADAPT, Wessel and Jurs (1995) developed a highly accurate linear seven-parameter model, with the correlation coefficient R^2= 0.994 and standard deviation s = 6.3K, to predict the normal boiling point of hydrocarbons. In a similar case of a more diverse set of 298 organic compounds, Katritzky et al. (1996) obtained a five-parameter model (R^2 = 0.9732, s = 12.41K) for prediction of boiling points using CODESSA. They claimed that the uncertainty in their predictions was equivalent to the experimental error (2.3% *vs.* 2.1%). Using the multivariate technique of partial least-squares (PLS) regression and a set of boiling points for 48 polycyclic aromatic compounds, Ferreira (2001) developed a relatively accurate 5-parameter model (R^2 = 0.9992 and s = 4.4K).

In a comparative QSPR study of the descriptors (topological indices versus electronic, geometrical, and combined molecular descriptors), Katritzky and Gordeeva (1993) reported six-parameter models of *BP*, density, *RI*, and other properties for a small data set of aldehydes (N ≤ 72), amines (N ≤ 110), and ketones (N ≤ 60). R^2 of their best BP models were 0.985, 0.982, and 0.991 for aldehydes, amines, and ketones, respectively. R^2 of their best density models were 0.941, 0.956, and 0.962 for aldehydes, amines, and ketones, respectively. R^2 of their *RI* models were 0.940, 0.954, and 0.985 for aldehydes, amines, and ketones, respectively. The five variables they used were selected from a set

of 84 descriptors. More descriptors (N > 450) can be accessed today through the CODESSA software.

In this chapter, we present a QSPR study of boiling points, densities, and *RIs* of pure hydrocarbons. The models were developed using the CODESSA software and two data sets: one including 186 saturates the other including 200 aromatics. Six eight-parameter multi-linear regression models were obtained for saturates and aromatics, separately. Using the combined saturate-aromatic data set, three additional models (for *BP*, SG, and *RI*) were developed for comparison. The validity and accuracy of these models were tested using the leave-one-out cross-validation method and data sets separate from those used for model fitting (Martens and Dardenne, 1998). The choice of descriptors for these models, based on a comparison of the best single-descriptor models, depended on the data set (saturates vs. aromatics) and on the property modeled (*BP*, SG, or *RI*). This study focuses on physical properties that were subject to relatively few previous QSPR studies, which were mostly devoted to a single property (Wessel and Jurs, 1995; Katritzky et al., 1996) or specific molecular type (Katritzky and Gordeeva, 1993). In addition, this study uses a relatively large data set compared to the studies mentioned above.

2.2. Data sets

Three data sets were used to develop the multi-linear regression models discussed below. Two separate sets included the boiling points, densities and refractive indices of saturate and aromatic hydrocarbons. The third set was a combination of the other two. The two original data sets contained the properties of 186 saturate and 200 aromatic hydrocarbons, respectively. In addition, separate smaller sets were prepared to validate

the models for individual properties after they were developed. The validation set for the saturated hydrocarbons simply included all the properties in question for 34 hydrocarbons. Three separate aromatics sets were prepared to validate the models for *BP* (61 hydrocarbons), *SG* (36 hydrocarbons), and *RI* (27 hydrocarbons). These test sets, used for the validation of the corresponding models, were not used for model fitting.

The saturate training set consisted of paraffins and cycloparaffins up to the middle distillate range. The aromatics training set contained mono-, di-, and tri-aromatics and included molecules containing naphthenic rings. Molecules larger or structurally different from those in the training sets were specifically chosen for test sets to assess the performance of our models in extrapolating to higher boiling points (where experimental data become scarce) and in distinguishing between differing structures within the same molecular mass (where the available group contribution methods fail). Values of normal *BP*, *SG* at 15.6°C, and *RI* at 25°C for all the 481 hydrocarbon compounds are tabulated in Appendix A together with their model predicted values.

The experimental data for our data sets were taken from the API technical data book (1995) and other sources (CRC Handbook of Chemistry and Physics, 2001; Bjørseth, 1983; Rossini, 1953; Karcher, 1988; Beilstein Crossfire Database, 2000). Most of the *SG* values were reported at 15.6°C. Those reported at 20°C were converted to 15.6°C using the density conversion tables for crude oils (ASTM D1250-80). Most of the *RI* data were reported at 25°C. Those reported *RI* at 20°C were converted to 25°C by applying a factor (0.9987) correlated from available *RI* at both temperatures. The accuracy of these conversions was checked using values known at both temperatures and found to be

reliable within ±0.001 for both *SG* and *RI*. Interpolation was used to correct data reported at different temperatures.

2.3. Methodology

2.3.1 Molecular descriptors

The QSPR analysis was conducted using commercial software, CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis). In CODESSA, a large number (>450) of molecular descriptors can be calculated (CODESSA Reference Manual, 1997). These descriptors can be divided into six groups: constitutional, topological, geometrical, electrostatic, quantum–chemical, and thermodynamic. The constitutional descriptors reflect the molecular composition of the compound without using its geometrical or electronic structures (e.g. number of atoms, molecular weight, number of rings, etc.). The topological descriptors describe the atomic connectivity in the molecule, such as Wiener index (Wiener, 1947), Randic indices (Randic, 1975), Kier shape indices (Kier, 1986), and Balaban index (Balaban, 1982). The Geometrical descriptors require 3-D coordinates of the atoms in the molecule (e.g. moments of inertia, shadow indices, molecular volume and surface descriptors). Based on the empirical partial charge calculations (Kirpichenok and Zefirov, 1987), the electrostatic descriptors describe the charge distribution of the molecule (e.g. minimum and maximum partial charge in the molecule, charged partial surface area, etc.). Hundreds of the quantum-chemical descriptors are calculated in CODESSA (Katritzky et al., 1997). They can be classified into five sub-groups: a) rigorous charge distribution descriptors; b) valence-related descriptors; c) quantum mechanical energy-related descriptors; d) molecular

rotational-vibrational descriptors; e) molecular solvation descriptors. The thermodynamic descriptors are derived from thermodynamic calculations using quantum mechanical techniques. More detailed description of the individual descriptors can be found in the CODESSA Reference Manual (1997).

2.3.2 Molecular modeling and input file preparation

The molecular 3-D structure of the molecule has to be optimized through energy-minimization before descriptor calculations. Because of their high efficiency, the semi-empirical quantum mechanical methods like PM3 and AM1 are widely used for molecular geometry optimization. The CAChE 5.04 software from Fujitsu with built-in MOPAC2002 and a graphical interface, was used for the PM3 molecular optimization because it gave the most accurate prediction in heat of formation for hydrocarbons (MOPAC2002 User's Manual, 2000). The final result of energy minimization calculations was the standard MOPAC output file accepted as the input file by CODESSA.

2.3.3 Descriptor screening and regression analysis

After the molecular modeling was completed, the MOPAC output files for individual compounds were loaded into the CODESSA program along with their corresponding physical properties (*BP*, *SG*, and *RI*). A set of 242 descriptors was then generated. Five additional descriptors – logarithm of Wiener index (LogW), square root MW ($MW^{1/2}$), cubic root of MW ($MW^{1/3}$), square root gravitational index (all bonds) $G^{1/2}$, and cubic

root gravitational index (all bonds) $G^{1/3}$ – were included because they were highly correlated to *BP* (Wessel and Jurs, 1995; Katritzky et al., 1996; Ferreira, 2001).

CODESSA includes two advanced procedures for systematic development of multi-linear QSPR equations: the *Heuristic* method and the *Best Multi-Linear Regression* (BMLR) method. The Heuristic method is usually used to pre-screen large data sets by eliminating highly correlated descriptors. As a result, a subset of the most important descriptors can be selected. The BMLR method offers a more systematic and thorough search of preferred descriptors. In this work, the BMLR analysis was employed to derive the best model because of its ease of implementation and the interpretability of the resulting equations (Stanton and Jurs, 1990). A pre-selection procedure was implemented that screened out some descriptors based on the intercorrelation coefficients in all possible descriptor pairs. For the pairs having the intercorrelation coefficient higher than 0.8 only one descriptor was included in the regression analysis. As a result, 76 descriptors were discarded and 166 were used in further analysis.

The search for the best QSPR model involves the generation of possible regressions and forward selection procedures. Detailed description of the method is available in the CODESSA Reference Manual (1997). In short, for the selected orthogonal pairs of descriptors, linear regression models were developed for the property in question with each descriptor. The descriptors that yielded high correlation coefficient R^2 values are used to perform the higher-order regression analysis. Fisher numbers are calculated for each model at 95% probability level. High order multi-linear models were selected if the Fisher criterion was larger than the one of lower order. At a specific order of regression, the best QSPR model was chosen based on the maximum value of Fisher number and

highest R^2. The selection of the final multilinear regression model was based on the Mallows' criterion (Neter et al., 1996) and standard deviation expectations. In general, the model is expressed as

$$PP = a_0 + \sum_{i=1}^{n} a_i x_i \tag{2-1}$$

where PP is the physical property in question, the parameter a_0 is the y-intercept, the parameter a_i is the coefficient by the ith descriptor x_i, and n is the number of descriptors in the final model.

2.3.4 Model validation

To check for overfitting, $R^2_{adjusted}$ and Mallows' Criterion (C_p) were calculated for the final models as follows (Neter et al., 1996)

$$R^2_{adjusted} = 1 - \frac{(1 - R^2)(n-1)}{n - p - 1} \tag{2-2}$$

where n is number of data set, p is number of parameters; and

$$C_p = \frac{SSE}{\sigma^2} - (n - 2p) \tag{2-3}$$

where SSE is the error of sum of square using p parameters; and σ is the estimation variance when full parameters are used. In estimation, σ can be obtained from the asymptotic value of standard deviation when increasing p. For example, $R^2_{adjusted}$ and C_p were calculated for three to eleven-parameter BP models for saturates with a cut-off point 0.0001 on R^2. $R^2_{adjusted}$ increases monotonously with p. The minimum C_p of 112, corresponding to the eleven-parameter model, was much larger than the number of parameters. Similar checks were carried out for all other models developed here and

overfitting was not observed. Therefore, it can be safely assumed that the eight-parameter model "standard" chosen in this paper for consistency would not lead to overfitting. Higher order QSPR models are found in the literature. For example, in their study of Charged Partial Surface Area (CPSA) descriptors, Stanton and Jurs (1990) developed a ten-parameter model for retention index using 107 data points and a twelve-parameter *BP* correlation using a 209 data set.

The leave-one-out cross-validation method (Martens and Dardenne, 1998) was used to validate the obtained QSPR models. For each experimental data point, the multi-linear regression was recalculated with the same descriptors but without this data point. Then the obtained regression equation was used to predict this data point. Finally, the obtained array of predicted data points was linearly correlated with the array of experimental data points, and the correlation coefficient, R^2_{CV}, was reported as the cross-validated correlation coefficient. Essentially, R^2_{CV} is a characteristic of the predictive power of the model. Smaller test sets, separate from the databases used in model development, were also used to check the predictive ability of all the models developed in this work. By design, the compounds included in this set had larger molecular size or were structurally different from those in the training set.

2.3.5. Inter-correlation of the modeled properties

To interpret and validate the QSPR models developed, it is important to assess inter-correlation between *BP*, *SG*, and *RI*. Table 2.1 shows such assessment for saturates, aromatics, and the combined set. Little intercorrelation was observed between *BP* and *SG* (R = 0.5165 for all hydrocarbons), and between *BP* and *RI* (R = 0.4922 for all

hydrocarbons). A higher inter-correlation was found between *SG* and *RI* (R = 0.9757 for all hydrocarbons). Similar conclusions were made by Katritzky and Gordeeva (1993) in their QSPR studies for hetero-compounds. This relationship is more significant for aromatics than that for saturates.

Table 2.1. Correlation coefficients between BP, SG, and *RI*

Property	*Saturates*			*Aromatics*			*Saturates + Aromatics*		
	BP	SG	*RI*	*BP*	SG	*RI*	*BP*	SG	*RI*
BP	1.0000	0.7131	0.6893	1.0000	0.4229	0.3974	1.0000	0.5165	0.4922
SG		1.0000	0.9611		1.0000	0.9837		1.0000	0.9757
RI			1.0000			1.0000			1.0000

2.4. Results and Discussion

2.4.1 Modeling of *BP* for saturates

Based on a cut-off point of 0.001 on R^2 improvement, a set of *BP* QSPR equations with 2 to 5 descriptors was obtained using the database of 186 structures. A three-parameter model (Eqn. 2-4) involving the number of rings (N_{ring}) and the cubic root of *MW* already gave a relatively high R^2 =0.9942 and s = ±11.42. The two descriptors are simple to calculate without aid of software. Unfortunately, such simple models were not obtained for the rest of the models of this work, owing to the inclusion of sophisticated descriptors in the first pair of variables selected.

$$BP_{sat} = -204.46 \pm 3.4268 + 16.017 \pm 0.9328 \times N_{ring} + 122.59 \pm 0.5779 \times \sqrt[3]{MW} \qquad (2\text{-}4)$$

Equation 2-4 is adequate for general use. However, with access to more sophisticated descriptors a five-parameter model is recommended (R^2 =0.9970 and s = ±8.30K).

$$BP_{sat} = -210.65 \pm 4.6092 + 15.717 \pm 0.7482 \times N_{ring} + 130.77 \pm 0.7312 \times \sqrt[3]{MW} - 3.5052 \times 10^{-3} \pm 2.5899 \times 10^{-4} \times WI - 6.5931 \pm 0.6772 \sqrt{Shadow_{YZ}} \qquad (2\text{-}5)$$

where *WI* is the Wiener Index and $Shadow_{YZ}$ is the YZ Shadow.

Since accurate molecular representation relies on accurate physical property estimation for pure compounds, an attempt was made in this work to derive highly accurate models with standard deviations close to the experimental errors in the database. Because the database uncertainties of *BP* for organic compounds were reported between 1% (Wessel and Jurs, 1995) and 2.1% (Egolf et al., 1994), corresponding to average errors of ±4.4K and ±9.0K, respectively, the *BP* model with five parameters was deemed sufficiently accurate for saturates. As shown below, the sufficiently accurate models for SG and *RI* required more parameters. The cut-off point of 0.001 on R^2 improvement generated eight- or nine-parameter, respectively. For consistency and comparison, eight-parameter models are used throughout this work.

The eight-parameter *BP* model for saturates produced R^2 = 0.9979 and s = ±6.10K, which was within the range of experimental uncertainties and corresponded well with other QSPR studies (Wessel and Jurs, 1995; Katritzky et al., 1996; Ferreira, 2001). This model is summarized in Table 2.2. Figure 2.1 shows the corresponding parity plot – the calculated versus experimental *BP*s. The variables include the number of rings, cubic root of MW, Wiener index, average information contents (order 0 and order 1), YZ shadow, and the polarity parameter. The number of rings represents the presence of cycloparaffinic structures. The cubic root of MW accounts for the non-linearity behavior of *BP* versus molecular size for saturates. Wessel and Jurs (1995) found that if the square root of MW were included in their model, the nonlinear behavior of the model would vanish. It is interesting to note that the cubic root of MW worked best in this work. The Wiener index and Average Information Contents (Shannon, 1948) appeared to quantify the effect of molecular topology that is well known to affect *BP*. *BPs* are also

significantly influenced by the dispersion forces, the strengths of which are determined by molecular shape reflected by the YZ shadow index - the orthogonal (YZ) projection of 3-D molecular shape (Rohrbaugh and Jurs, 1987). The polar compounds have higher BP than non-polar at the same molecular size, due to the stronger intermolecular forces involved. The polarity parameter in our model, the difference between maximum partial charge and minimum partial charge of the atoms, accounts for this effect.

Table 2.2. BP model for saturates

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	-3.1747E+02	1.5723E+01	-20.1908	Intercept
1	1.5237E+01	6.3842E-01	23.8664	Number of rings
2	1.2693E+02	7.5909E-01	167.2206	Cubic root of MW
3	-4.3372E-01	4.9485E-02	-8.7648	YZ Shadow
4	-3.2707E-03	2.2473E-04	-14.5539	Wiener index
5	8.7891E+01	8.3751E+00	10.4943	Average Information content (order 0)
6	-3.6653E+01	3.6166E+00	-10.1348	Average Information content (order 1)
7	1.4701E+03	3.4850E+02	4.2183	Polarity parameter (Qmax-Qmin)

R^2 = 0.9979, s = ±5.87K, F = 13332.1, N = 186 compounds

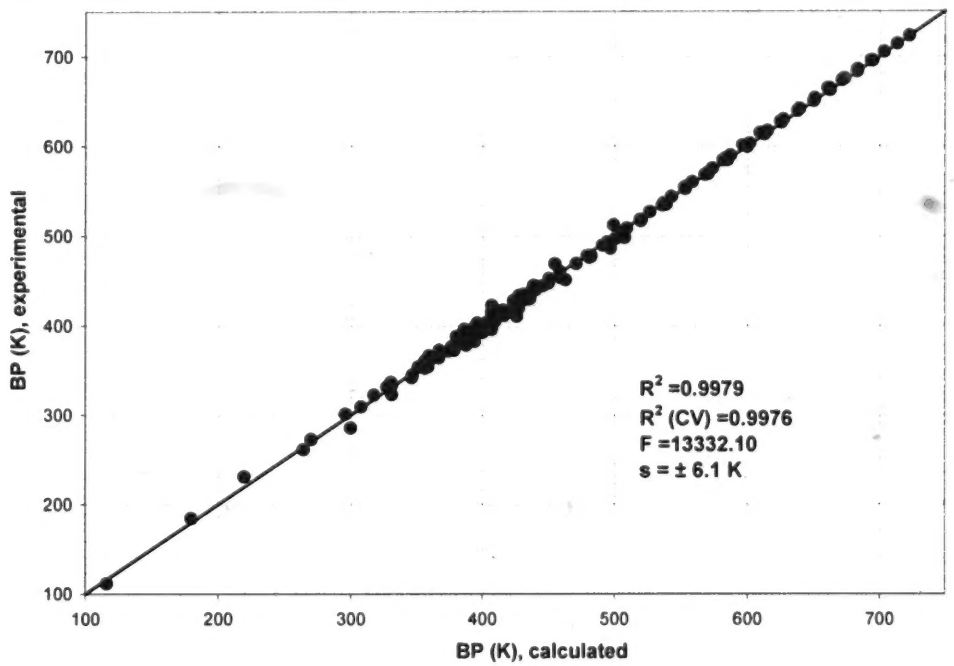


Figure 2.1. Parity plot of the BP model for saturates; 186-point training set

The cross-validation procedure applied to this model resulted in $R^2_{CV} = 0.9976$, which was practically the same as the model $R^2 = 0.9979$, indicating a good quality model. It was also reassuring to find that the *BP* of 2,2,4,4-tetramethyl-pentane, treated as an outlier in Wessel and Jurs' QSPR models (1995), was predicted by our model with a better accuracy (s = 11.46K). Figure 2.2 shows the parity of *BPs* calculated with our model with the experimental data in the separate test set, and compares it with *BP*s calculated using Joback's group contribution model (Joback, 1984). Our model predicts *BP*s of saturates well, with the average error less than 0.7% and standard deviation s = ±6.40K, while the predictions of the group contribution method were significantly worse, with half of the values outside of the plot area.

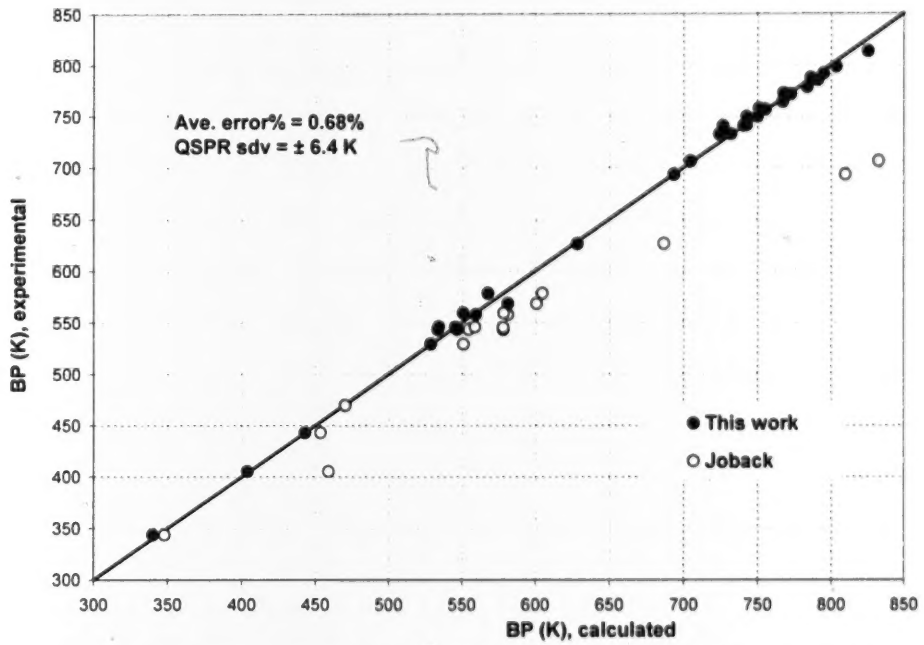


Figure 2.2. Comparison of the BP model for saturates with Joback's method
Parity plot for the 34-point test set

2.4.2 Modeling of *BP* for aromatics

This training set consisted of 200 mono-, di-, and tri-aromatic compounds, including aromatics with naphthenic rings. The eight-parameter multi-linear *BP* model obtained is summarized in Table 2.3. The R^2 of the model was 0.9960, and the standard deviation was ±6.30K. The corresponding parity plot is shown in Figure 2.3.

Table 2.3. BP model for aromatics

i	a_i	$\pm\Delta a_i$	*t*-test	x_i
0	8.9416E+01	1.1961E+01	7.4755	Intercept
1	9.1259E+03	5.6001E+02	16.2959	FNSA-3 Fractional PNSA (PNSA-3/TMSA) [Zefirov's PC]
2	1.7021E+01	4.3222E-01	39.3809	Square root of Gravitational Index (all bonds)
3	2.1177E+01	1.7073E+00	12.4038	Number of benzene rings
4	6.2893E+01	3.9505E+00	15.9202	SA-2 Fractional PPSA (PPSA-2/TMSA) [Quantum-PC]
5	1.1130E+00	7.3684E-02	15.1049	YZ Shadow
6	7.4707E+01	1.0525E+01	-7.0981	Average Structural Information content (order 1)
7	1.8355E+01	2.2858E+00	8.0301	Total point-charge component of the molecular dipole

R^2 = 0.996, s = ±6.2K, F = 6351.6, N = 200 compounds

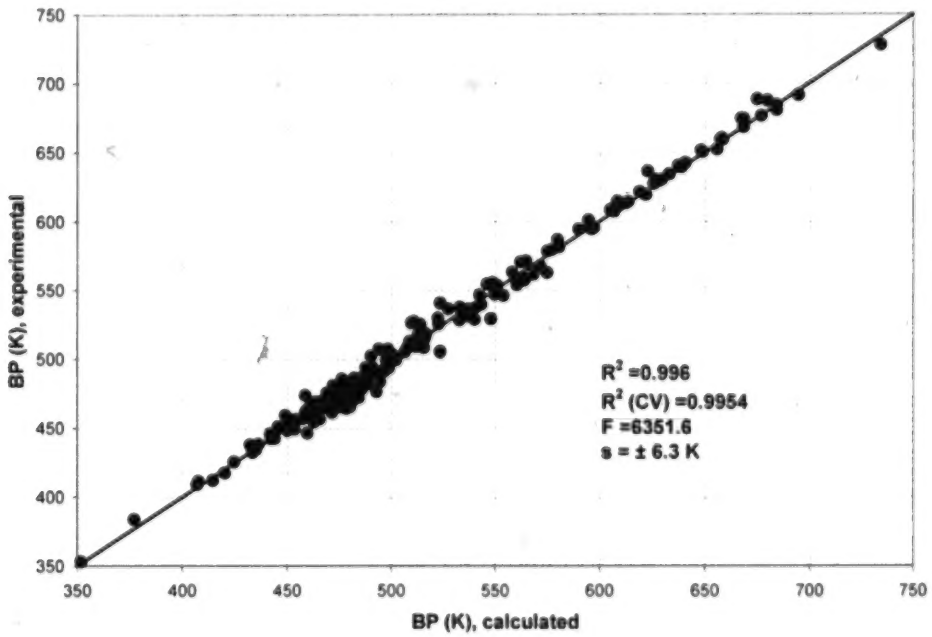


Figure 2.3. Parity plot of the BP model for aromatics; 200-point training set

Among the 7 descriptors, the Number of Benzene Rings was an obvious choice for aromatics. In a similar QSPR study, Katritzky et al (1996) found that the square root of the gravitational index, G_I, gave the highest R^2 in a one-variable model; but the cubic root of the gravitational index contributed most to the models when combined with the other parameters. In this study, the best result was obtained when the $G_I^{1/2}$ was used. Aromatic compounds are usually more polar than saturates. The polar intermolecular interactions were well accounted for by the Charged Partial Surface Area (CPSA) descriptors, which combine the solvent-accessible surface areas with partial atomic charges (Stanton and Jurs, 1990). Jurs and co-workers successfully applied these descriptors in the QSPR models for *BP* (Wessel and Jurs, 1995; Egolf et al., 1994). The shape (YZ shadow) and the structural information content descriptors were also included in the *BP* model for aromatics. The fact that these descriptors were involved in both saturate and aromatic models indicates that the molecular size, shape, and topological descriptors are important in BP prediction regardless of molecular classes in contribution to the physical properties.

The leave-one-out cross validation resulted in R^2_{CV} = 0.9954, only slightly smaller than that of this model. A test data set containing 61 aromatic compounds with up to 7 aromatic rings was used to further validate this model. The predicted BP results compared with the corresponding experimental values in Figure 2.4 indicate excellent predictive ability of the model with s = 7.24K. Only one compound, o-terphenyl, was flagged as an outlier. The reported *BP* of o-terphenyl was about 40°C less than that of m-terphenyl or p-terphenyl. Apparently, our model cannot predict differences of this magnitude. This compound was also marked as an outlier by Wessel and Jurs (1995).

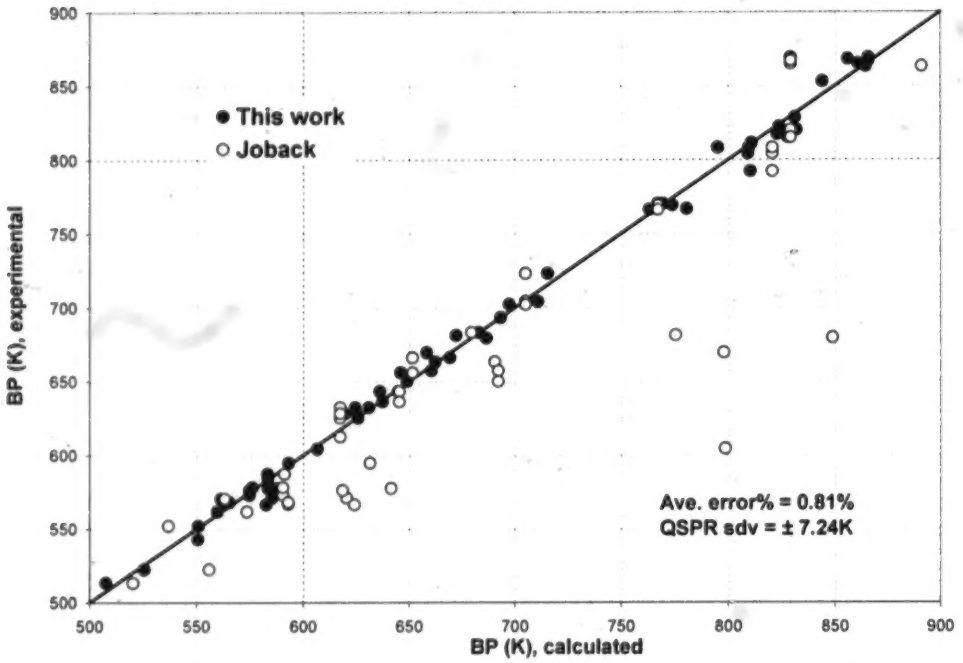


Figure 2.4. Comparison of the BP model for aromatics with Joback's method
Parity plot for the 61-point test set

Figure 2.4 also compares predictions of our model with those of Joback's group contribution method. Joback's method (1984) did not work well for the large molecules included in the test set as indicated by the large scatter, with approximately 1/3 of the points found outside of the plot area.

2.4.3 Modeling of *BP* for both saturates and aromatics

As part of this study, we explored the idea of creating a single model covering both saturates and aromatics. Table 2.4 summarizes an eight-parameter model developed using the combined, saturate and aromatic, training data set of 386 data points. Among the 7 variables, the three constitutional descriptors, number of rings, cubic root of MW, and number of single bonds reflected the characteristics of saturates and alkyl substituents on aromatic cores. Three CPSA descriptors (FPSA-3, PNSA-3, and FPSA-3) selected in the

model reflected the importance of polar interactions between molecules. The parity plot for this model and the combined data set is shown in Figure 2.5. As expected, R^2 = 0.9947 was lower than those obtained for saturates (R^2 = 0.9979) or aromatics (R^2 = 0.9960) separately. Similarly, s = 9.40K was greater than s = 6.10K for saturates or s = 6.30K for aromatics. Also, the leave-one-out cross validation of this model gave a slightly lower R^2_{CV} = 0.9938.

Table 2.4. BP model for both saturates and aromatics

i	a_i	$\pm\Delta a_i$	*t*-test	x_i
0	-1.7300E+02	11.018	-15.7014	Intercept
1	2.7537E+01	0.70061	39.304	Number of rings
2	1.6884E+02	2.2757	74.194	Cubic root of MW
3	-1.4379E+00	0.07119	-20.1986	YZ Shadow
4	-1.6471E+00	0.082867	-19.8761	Number of single bonds
5	-1.1434E+04	883.33	-12.944	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Zefirov's PC]
6	1.5092E+01	1.5666	9.6331	PNSA-3 Atomic charge weighted PNSA [Zefirov's PC]
7	-6.1426E+02	88.580	-6.9346	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Quant-PC]

R^2 = 0.9947, s = ±9.4K, F = 12668.1, N = 386 compounds

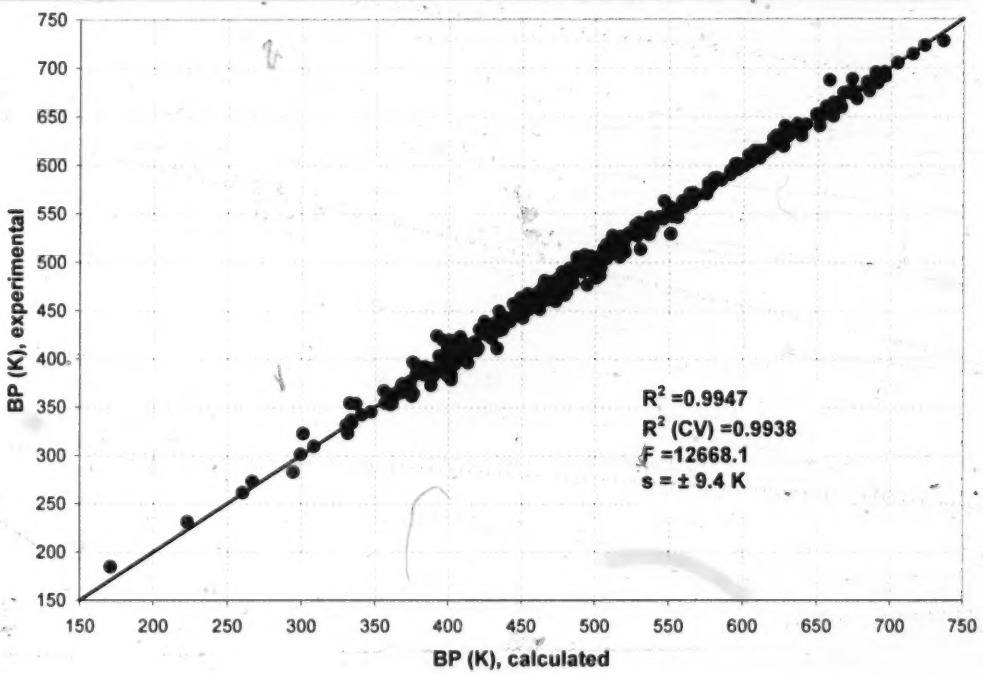


Figure 2.5. Parity plot of the BP model for saturates and aromatics; 386-point training set

2.4.4 Modeling of *SG* for saturates

It is thought that density is somewhat related to the BP of chemical compounds. For instance, Retzekas et al. (2002) recently used density as one of the parameters to correct the prediction errors in *BP* of Joback's method. In the large database we compiled this is not the case (see Table 2.1). Consequently, using this database, the descriptors selected for the *SG* model would be expected to differ from those for the *BP* model. Using the 186-point saturate training set, an eight-parameter model for *SG* with $R^2 = 0.9910$ and s = 0.007 was developed. The model is summarized in Table 2.5. Smittenberg and Mulder (1948) reported that the densities of n-paraffins and n-alkyl ring compounds change linearly with the reciprocal of MW. However, the relative molecular weight (the molecular weight divided by number of atoms) was chosen as a variable in our model instead. For hydrocarbons, the relative molecular weight is a function of H/C ratio. The Balaban index and YZ shadow characterizes the molecular complexity and shape, which directly influence the compactability of molecules. In contrast to the *BP* model for saturates, three CPSA descriptors (RPCG, RNCS, and FPSA-3) were selected for the *SG* model. This suggests that the polar intermolecular interactions play an important role in *SG* estimation for saturates.

Table 2.5. *SG* model for saturates

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	-7.1069E-01	3.2158E-02	-22.1002	Intercept
1	2.6370E-01	8.0553E-03	32.7366	Relative molecular weight
2	1.3390E-02	1.5649E-03	8.5567	Balaban index
3	-1.0662E+00	4.4698E-02	-23.8547	RPCG Relative positive charge (QMPOS/QTPLUS)
4	1.1329E-02	6.5671E-04	17.2517	RNCS Relative negative charged SA (SAMNEG*RNCG)
5	1.4926E-03	1.5995E-04	9.3317	YZ Shadow
6	4.2495E+00	4.8341E-01	8.7908	FPSA-3 Fractional PPSA (PPSA-3/TMSA)
7	5.0865E+00	7.6447E-01	6.6536	Min partial charge for a H atom [Zefirov's PC]

$R^2 = 0.9910$, $s = \pm 0.007$, $F = 1925.44$, $N = 186$ compounds

The parity plot for the *SG* model is shown in Figure 2.6. The calculated *SGs* agree well with the experimental values of the training saturate set. The cross-validation R^2_{CV} = 0.9894 was slightly lower than that of the *SG* model (R^2 = 0.9910). The parity plot comparing the *SG* model and 34-point saturate test sets is given in Figure 2.7. The average relative prediction error was 1%. Compared to that of the training set, s increased from 0.007 to 0.01. Densities estimated from the empirical formula of Rackett (1970) are also plotted in Figure 2.7. Rackett's formula uses the critical properties (P_c, T_c, V_c) to calculate density of saturated organic liquids. In some cases, where the experimental data were not available, the critical properties were calculated using Joback's method (Joback, 1984). Rackett's predictions are highly scattered with more than half of the calculated densities falling outside of the plot.

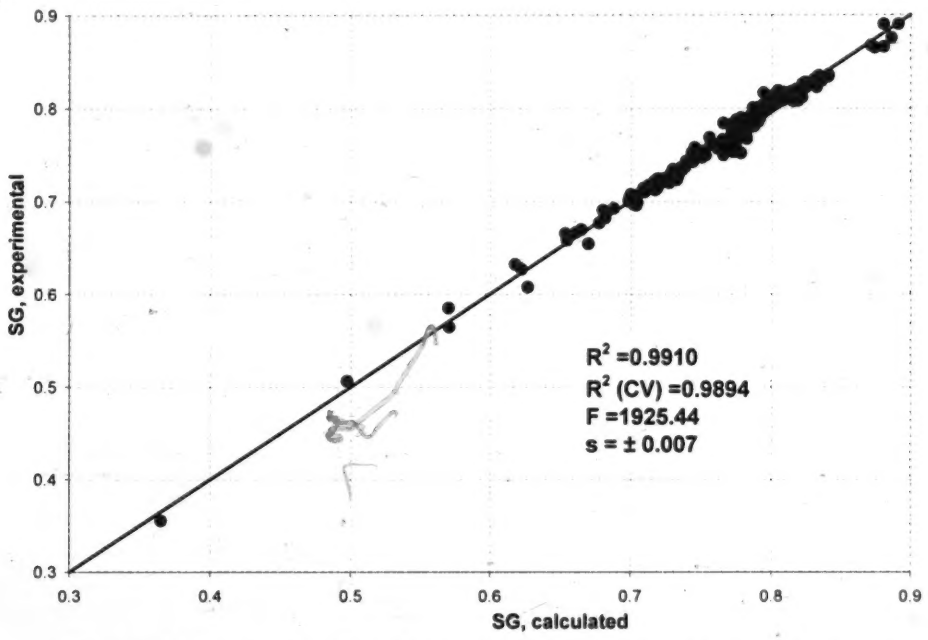


Figure 2.6. Parity plot of the SG model for saturates; 186-point training set

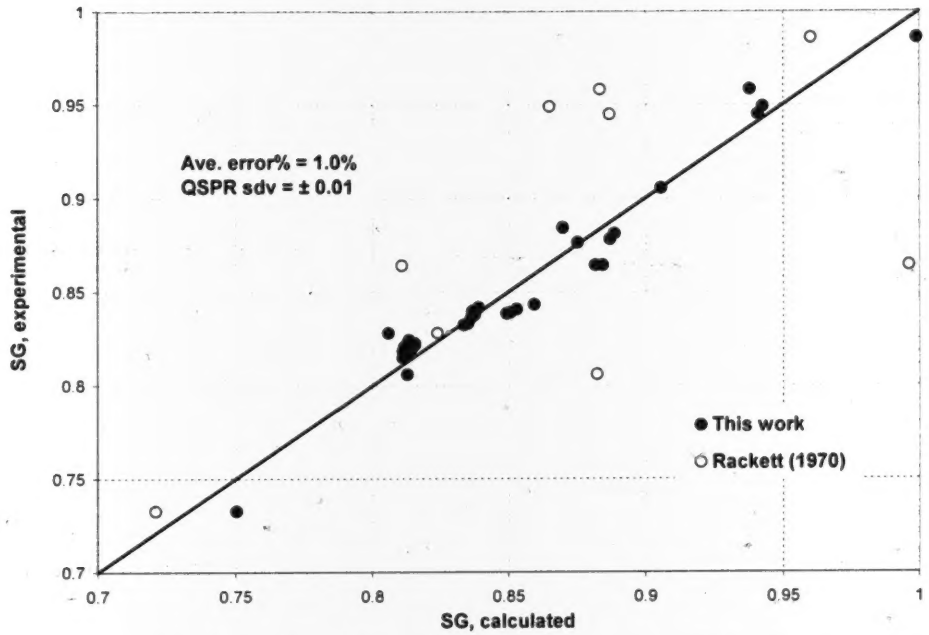


Figure 2.7. Comparison of the SG model for saturates with Rackett's method
Parity plot for the 34-point test set

2.4.5 Modeling of *SG* for aromatics

The best multi-linear regression model for *SG* of aromatics was based on the same aromatic training set used for BP modeling. The resulting eight-parameter model (R^2 = 0.9881, s = 0.008) is summarized in Table 2.6. The set of variables in this model includes two constitutional indices (Relative Number of Rings, and $G^{1/3}$), three topological indices (Balaban index, Topographic Electronic index, and Average Information Content), one geometrical index (Moment of Inertia C, which describes the molecular mass distribution relative to Z axis), and one quantum-chemical descriptor (Total Point-Charge Component of Molecular Dipole). The number of rings affects the aromatics density in a reduced form — the relative number of rings (the number of rings divided by the number of atoms). It is interesting to note that, in combination with the other six descriptors, $G^{1/3}$

gave the best model for *SG* of aromatics, while in the *BP* model for aromatics $G^{1/2}$ gave the best result. The selected three topological indices indicate a significant effect of molecular structural complexity on density.

Table 2.6. *SG* model for aromatics

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	2.4016E-01	2.6735E-02	8.9832	Intercept
1	2.0875E+00	4.7869E-02	43.6081	Relative number of rings
2	5.7249E-02	2.3020E-03	24.869	Cubic root of Gravitational Index (all bonds)
3	-1.1601E-01	6.1036E-03	-19.0074	Topographic electronic index (all bonds) [Zefirov's PC]
4	4.5907E-02	2.4801E-03	18.5102	Balaban index
5	3.4568E-02	4.3835E-03	7.8859	Total point-charge component of the molecular dipole
6	-2.4902E-02	3.4860E-03	-7.1433	Average Information content (order 1)
7	-9.4386E-01	1.3804E-01	-6.8378	Moment of inertia C
R^2 = 0.9881, s = ±0.008, F = 1820.91, N = 200 compounds				

Figure 2.8 shows the parity plot for the *SG* model for aromatics. Although R^2= 0.9881 for the eight-parameter model was not as high as that for the aromatics *BP* model (R^2 = 0.9960), our intermediate six-parameter model with R^2 = 0.9861 was already superior to the published QSPR density models (Katritzky and Gordeeva, 1993), where the R^2s of their 5-variable models were less than 0.96. R^2_{CV} = 0.9863 was slightly smaller than that of the model itself. This model was also used to predict *SG* for the aromatics test set containing 36 data points. The predicted *SGs* plotted in Figure 2.9 were consistent with the experimental data, except for 9,10-dihydro-anthracene. The reported densities of this compound were 1.215 g/cm^3 from the CRC handbook, and 0.88 g/cm^3 from the Aldrich chemical catalog (both at 20°C). The predicted value was 1.0977. Therefore, this compound was treated as an outlier. The average prediction error for the test set was 0.8% and s increased from 0.008 for the training set to 0.010 for the test set. A comparison between the highly scattered densities estimated using Rackett's method (1970) and the measured ones is also given in Figure 2.9.

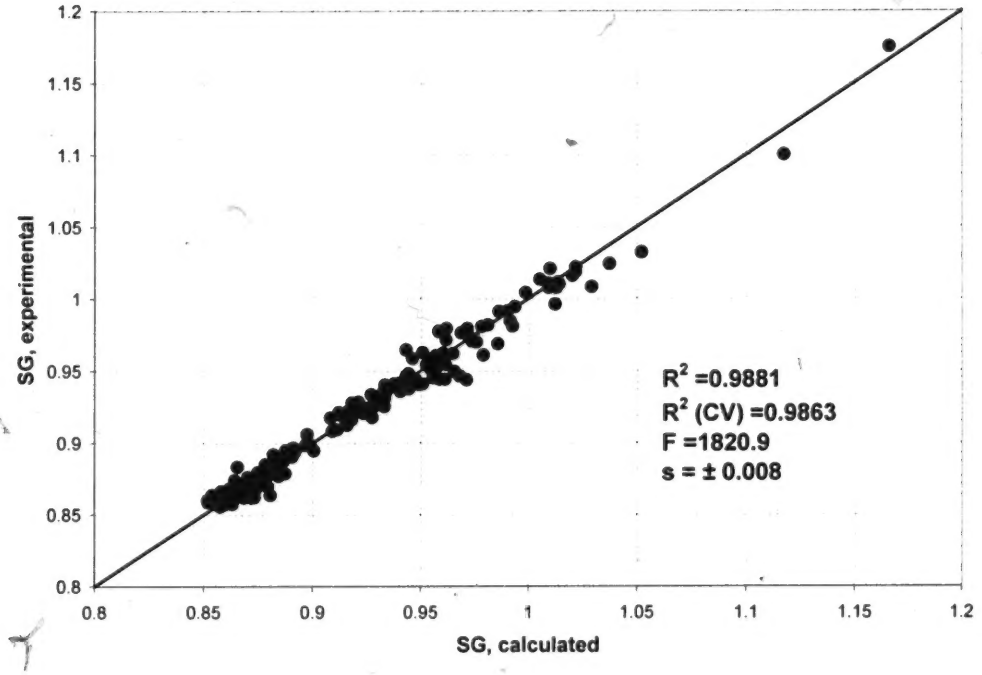


Figure 2.8. Parity plot of the SG model for aromatics; 200-point training set

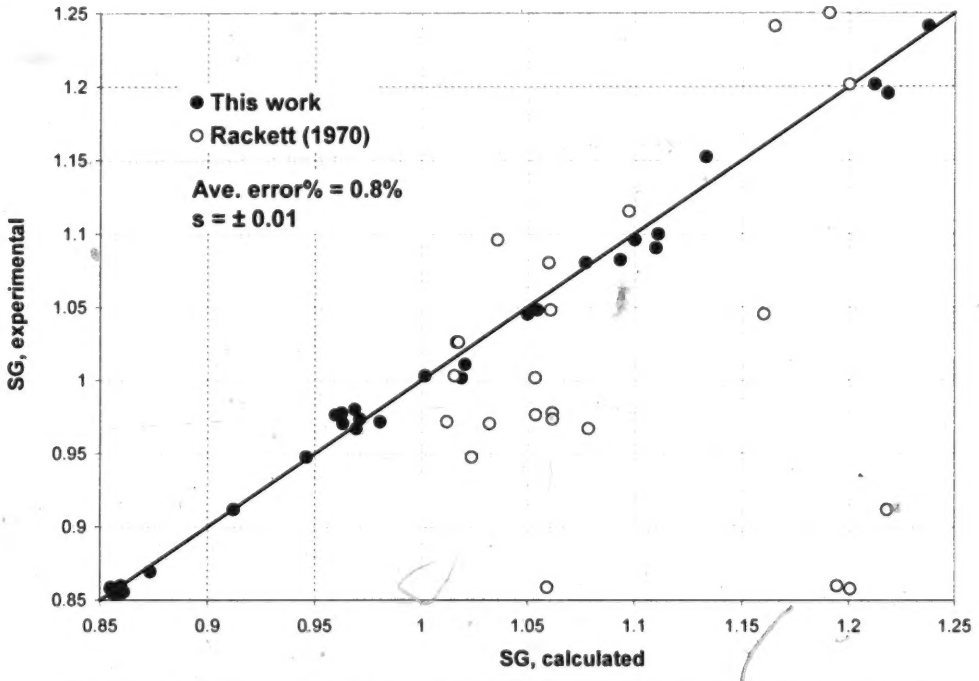


Figure 2.9. Comparison of the SG model for aromatics with Rackett's method
Parity plot for the 36-point test set

2.4.6 Modeling of *SG* for both saturates and aromatics

The eight-parameter multi-linear regression *SG* model for the combined 386-point training set, including both saturates and aromatics, is summarized in Table 2.7. Five of the seven descriptors (Relative Number of Rings, $G^{1/3}$, Balaban Index, Topological Electronic Index, as well as the Total Point-Charge Component of the Molecular Dipole) were the same as those in the *SG* model for aromatics, which reiterated their general importance as factors affecting density.

Table 2.7. *SG* for both saturates and aromatics

i	a_i	$\pm\Delta a_i$	*t*-test	x_i
0	-3.5119E-03	1.5648E-02	-0.2244	Intercept
1	1.6851E+00	5.7082E-02	29.5209	Relative number of rings
2	7.3419E-02	1.9256E-03	38.1275	Cubic root of Gravitational index (all bonds)
3	-1.3017E-01	6.9024E-03	-18.8583	Topographic electronic index (all bonds)
4	2.9772E-02	1.6360E-03	18.1976	Balaban index
5	1.9088E-02	2.6065E-03	7.323	Total point-charge component of molecular dipole
6	-4.6003E-03	5.7516E-04	-7.9982	Bonding Information content (order 1)
7	7.2553E+00	1.1138E+0	6.5139	Max partial charge for a H atom [Zefirov's PC]
R^2 = 0.9805, s = ±0.017, F = 2974.24, N = 386 compounds				

The parity plot for the *SG* model predictions and the combined data set of saturates and aromatics are given in Figure 2.10. Although a higher Fisher number was obtained due to a larger data set, R^2 dropped to 0.9805 compared to that for aromatics alone (R^2 = 0.9881). The cross-validation gave a slightly lower R^2 =0.9761, indicating good stability of the model. However, the model standard deviation increased significantly from ±0.007/0.008 for saturates/aromatics alone to 0.017 for both. For better accuracy, the use of separate models for saturates or aromatics is recommended

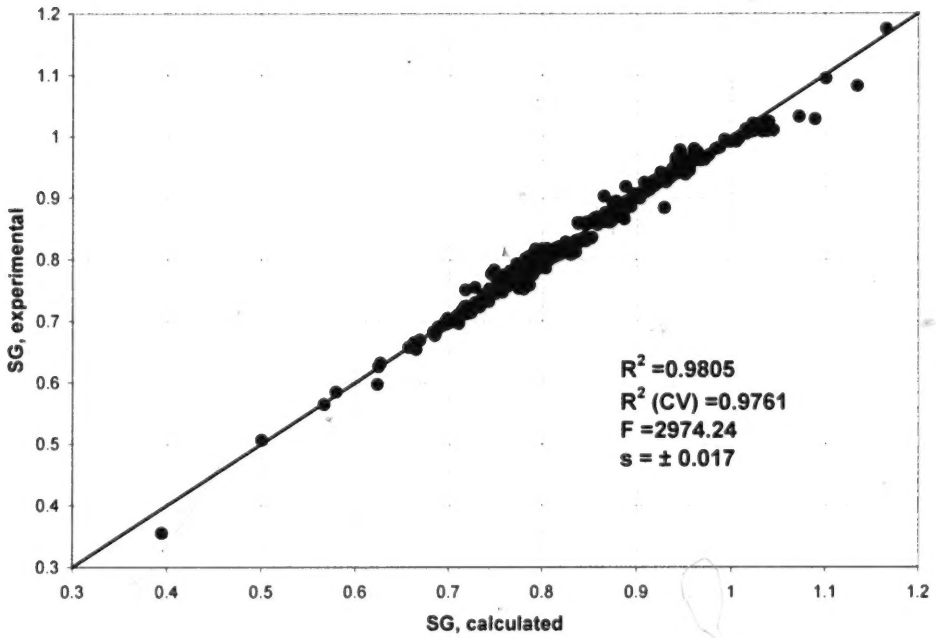


Figure 2.10. Parity plot of SG model for saturates and aromatics; 386-point training set

2.4.7 Modeling of *RI* for saturates

The eight-parameter model for *RI* is summarized in Table 2.8. Compared to the *SG* model for saturates, the *RI* model is slightly better with $R^2 = 0.9921$ (vs. $R^2 = 0.9910$) and $s = 0.004$ (vs. s = 0.007). As shown in Table 2.1, the study of the property model reveals that the refractive index is highly correlated with *SG* but has no significant relationship with *BP*. It is not surprising that *RI* correlated in a manner somewhat similar to density, but different from *BP*. Among the seven variables of this model, six (Relative Number of C Atoms, Balaban Index, Minimum Partial Charge, and three CPSA descriptors) were similar to those of the *SG* model. The Relative Number of C Atoms played a similar role to the Relative *MW* in the *SG* model. The CPSA and Partial Charge descriptors captured the effect of polar intermolecular interactions on *RI*. The Randic index, capturing the

contribution of the atom connectivity on *RI*, was the only descriptor significantly different from the *SG* model for saturates.

Table 2.8. *RI* model for saturates

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	1.0214E+0	1.5762E-02	64.8004	Intercept
1	8.9496E-01	4.5012E-02	19.8827	Relative number of C atoms
2	-4.4050E-5	4.5489E-06	-9.6837	WPSA-2 Weighted PPSA (PPSA2*TMSA/1000)
3	-2.6897E-1	8.8952E-03	-30.2383	RNCG Relative negative charge (QMNEG/QTMINUS)
4	7.0284E-03	5.9232E-04	11.866	Randic index (order 3)
5	1.9006E+0	2.3561E-01	8.0669	FPSA-3 Fractional PPSA (PPSA-3/TMSA) [Quantum-PC]
6	6.4486E-03	7.5071E-04	8.5901	Balaban index
7	2.1175E+0	3.6267E-01	5.8386	Min partial charge for a H atom [Zefirov's PC]
		R^2 = 0.9921, s = ±0.004, F = 3054.28, N = 186 compounds		

A parity plot of calculated versus experimental *RI* data for saturates is shown in Figure 2.11 The cross-validation check generated an only slightly lower correlation coefficient (R^2_{CV} =0.9902) confirming the validity of the model. Furthermore, a comparison of the model with the 34-compound test set (see Figure 2.12) confirmed its predictive ability. The average prediction error was 0.4% and s= ±0.006 slightly increased from s= 0.004 for the model. Again, the saturate test set included larger and more complex saturated hydrocarbons only, resulting in a relatively narrow range of data (*RI* = 1.40-1.52) in Figure 2.12. Although the model predictions were not as accurate as the experiments (ASTM D1218 repeatability = 0.0002), our model provided better estimates than other reported models. In a similar study (Katritzky and Gordeeva, 1993), for aldehydes (n = 60), amines (n = 110), and ketones (n = 59), the highest R^2 reported in their six-parameter model was R^2 = 0.9400. A significantly higher R^2 = 0.9883 was found in this work when we built a six-parameter model.

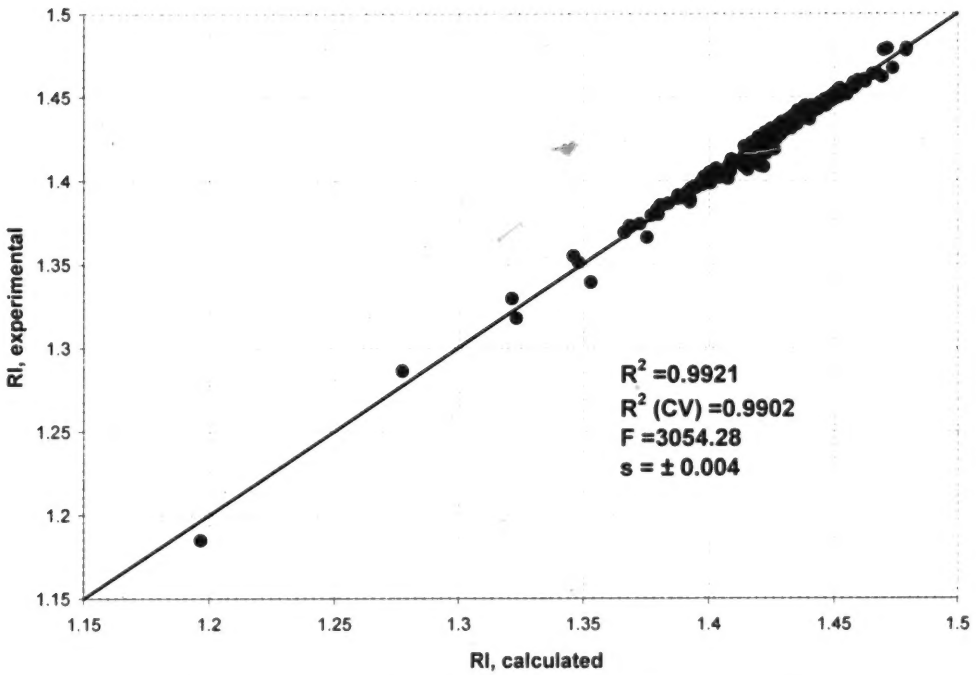


Figure 2.11. Parity plot of the *RI* model for saturates; 186-point training set

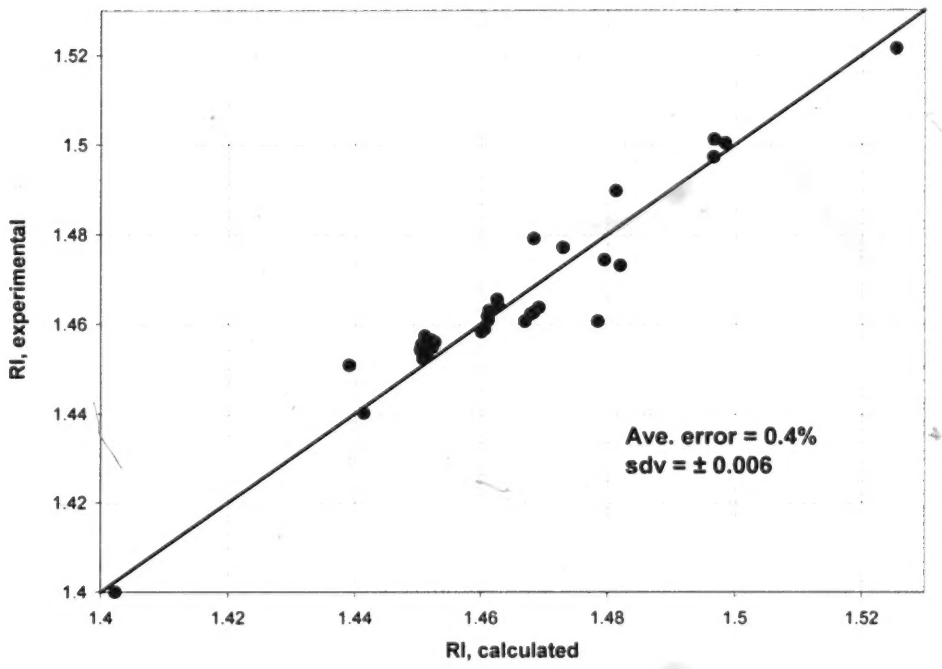


Figure 2.12. Parity plot of the *RI* model for saturates with the 34-point test set

2.4.8 Modeling of *RI* for aromatics

The best eight-parameter model for *RI,* based on the 200-compound aromatics training set, is summarized in Table 2.9. Like in all the models for *SG* and *RI* described above, the constitutional descriptors and topological indices were the most important descriptors. The ring structures played an important role in estimating the *RI* of aromatics. Both the Relative Number of Benzene Rings and the Number of Rings (aromatic and naphthenic) were included in the model. The Balaban index and logarithm of the Wiener index were chosen to represent the contribution of atomic connectivity on *RI*. The Wiener index was found to perform well in *BP* models (see Table 2 for saturates). It is worth noting that the Wiener index gave the best fit for *RI* in a reduced form (logarithm of Wiener index). Similar to the *SG* model of aromatics, the remaining three descriptors were selected from geometrical and partial-charge related descriptors (YZ shadow, Total Point-Charge Component of the Molecular Dipole, and Total Charged Weighted PPSA). A further comparison of these three descriptors in both models suggests that geometric descriptors were more important in the *SG* model while the polar intermolecular interactions were more important for the *RI* model for aromatics.

Table 2.9. *RI* model for aromatics

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	1.1340E+0	1.1377E-02	99.6729	Intercept
1	1.1154E+0	3.8967E-02	28.6235	Relative number of benzene rings
2	3.2522E-02	1.5726E-03	20.681	Log Wiener Index
3	5.0363E-02	1.3357E-03	37.7056	Number of rings
4	4.8150E-02	2.2964E-03	20.9673	Balaban index
5	-1.1022E-3	9.3164E-05	-11.8303	YZ Shadow
6	-6.1726E-5	7.3306E-06	-8.4203	PPSA-2 Total charge weighted PPSA [Zefirov's PC]
7	1.7142E-02	2.5094E-03	6.831	Total point-charge component of molecular dipole
R^2 = 0.9902, s = ±0.005, F = 2052.26, N = 200 compounds				

The parity plot between the calculated observed *RIs* is shown in Figure 2.13 for the 200-aromatic training set. Although the obtained *RI* model was not as good as the one for saturates, a relatively high R^2 =0.9902 was obtained, with a standard deviation of ±0.005. The cross-validation check on the model produced a slightly lower R^2_{CV} =0.9881. The model was also tested against the test set containing 27 aromatic compounds (see Figure 2.14). The average prediction error was 0.53% with an increased standard deviation ±0.01 (vs. 0.005 for the training). The increased prediction errors may have partially to do with increased uncertainties in experimental data for large hydrocarbons. At room temperature, larger molecules like naphthalene crystallize and the direct *RI* measurements at liquid phase are not possible. Considering the higher uncertainty for larger aromatic compounds, our model predictions are satisfactory.

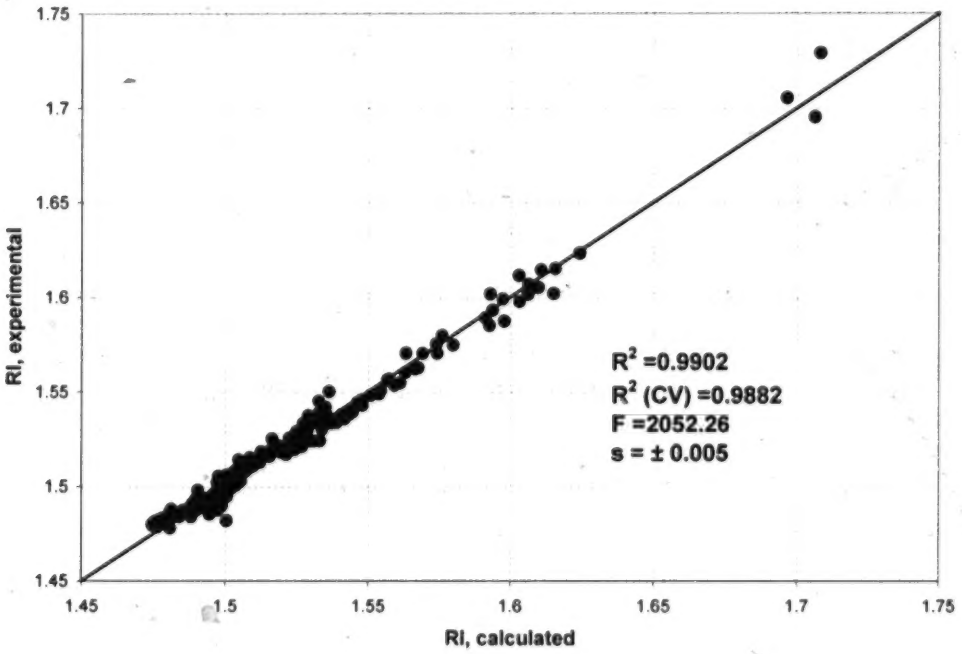


Figure 2.13. Parity plot of the *RI* model for aromatics; 200-point training set

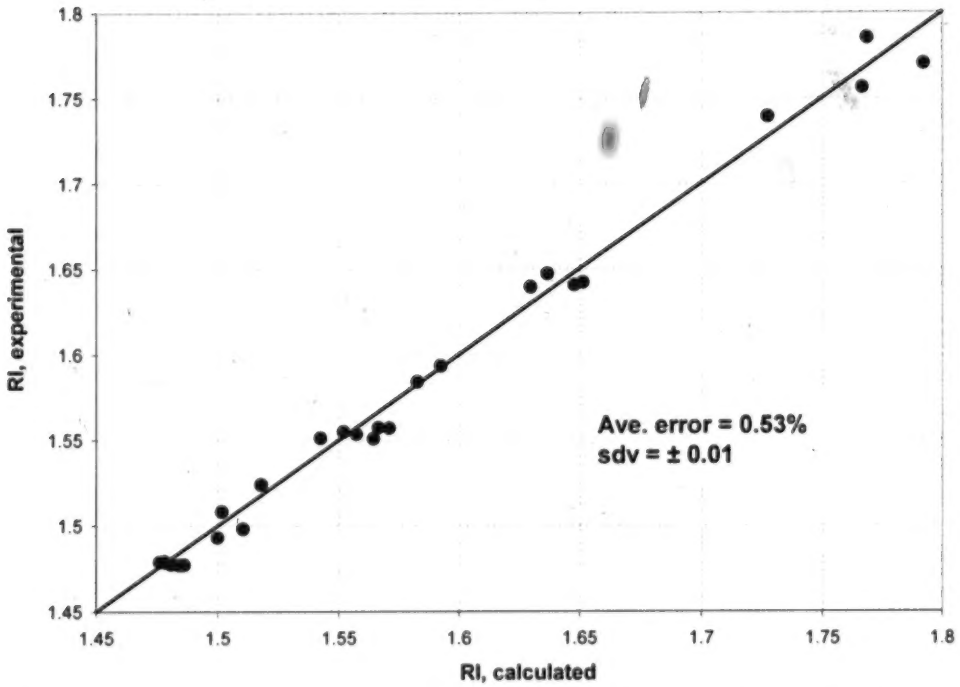


Figure 2.14. Parity plot of the *RI* model for aromatics with the 27-point test set

2.4.9 Modeling of *RI* for both saturates and aromatics

The best 8-parameter model for the combined saturate and aromatic training data set is summarized in Table 2.10. The selected descriptors reflected the characteristics of both saturates and aromatics in the model for *RI*. Among the 7 variables, four were exactly the same as those in the aromatics model (Relative Number of Benzene Rings, Balaban index, Number of Rings, and Total Point-Charge Component of the Molecular Dipole). The effect of the polar intermolecular interactions was captured through four partial charge related descriptors (RNCG Relative Negative Charge, PNSA-3, Total Point-Charge Component of the Molecular Dipole, and RPCG Relative Positive Charge). These results indicate that the ring structure, molecular topology, and polarity were the key factors influencing the refractive index of pure hydrocarbons in general.

Table 2.10. *RI* model for both saturates and aromatics

i	a_i	$\pm\Delta a_i$	t-test	x_i
0	1.4301	3.3390E-03	428.2983	Intercept
1	2.0788	5.0166E-02	41.4391	Relative number of benzene rings
2	-0.32089	1.7701E-02	-18.1281	RNCG Relative negative charge (QMNEG/QTMINUS)
3	3.4361E-2	1.1279E-03	30.4659	Number of rings
4	1.5248E-2	1.0170E-03	14.9934	Balaban index
5	2.9608E-3	2.4057E-04	12.3071	PNSA-3 Atomic charge weighted PNSA [Quantum- PC]
6	2.5519E-2	1.5373E-03	16.6006	Total point-charge component of the molecular dipole
7	-0.42067	4.8361E-02	-8.6985	RPCG Relative positive charge (QMPOS/QTPLUS)
		R^2 = 0.9881, s = ±0.008, F = 4774.81, N = 386 compounds		

The parity plot of calculated *versus* observed *RIs* is shown in Figure 2.15. R^2=0.9881 was lower than that for saturates (0.9921) or for aromatics (0.9902). The cross-validation check gave a slightly lower R^2_{CV} = 0.9854. Compared to the models for saturates and aromatics, the combined model was less accurate. The standard deviation for *RI* increased to 0.008 from 0.004 for saturates and 0.005 for aromatics. Again, for better accuracy, the use of separate models for saturates or aromatics is recommended.

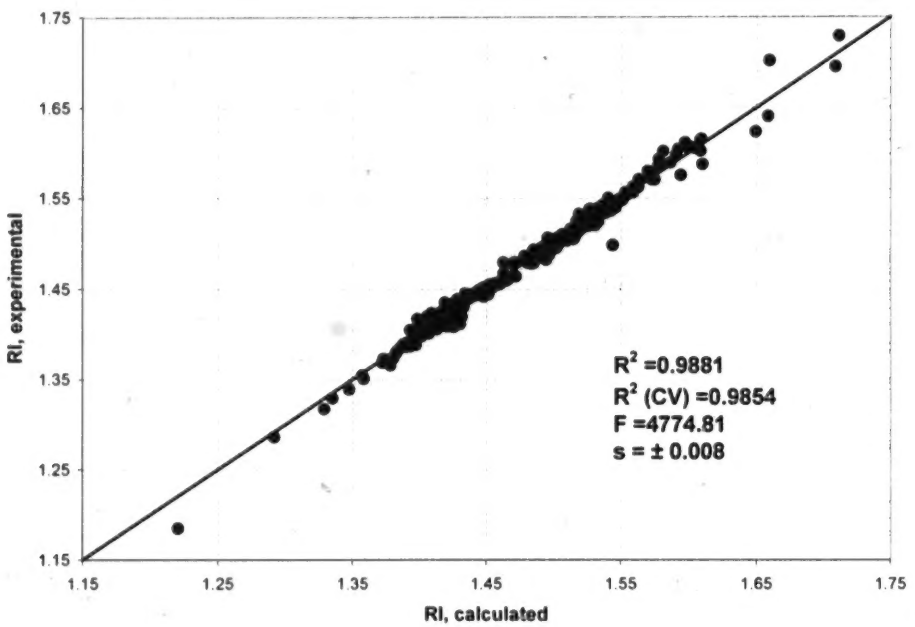


Figure 2.15. Parity plot of *RI* model for saturates and aromatics; 386-point training set

2.5 Conclusions

This study focused on the development of QSPR models for *BP*s, *SG*s, and *RI*s of pure hydrocarbons to enable estimation of bulk physical properties of mixtures of molecules to represent refinery streams. In our opinion, the developed multi-linear models are sufficiently accurate for this task and, at this time, there is no need to develop more sophisticated nonlinear models. On one hand, the properties selected for this study are easy to generate experimentally for those streams. On the other, relatively simple mixing rules can be used to estimate these properties for the molecular representations of those streams. As a result, this study enabled us to validate the choices of molecules for the molecular representations through the estimation of bulk properties of mixtures of molecules.

A set of nine eight-parameter models was developed in this study using the automated procedure of the CODESSA software. Except for the *SG* model for aromatics (R^2 = 0.9881), the correlation coefficient R^2 for all the models developed from separate saturates and aromatics were > 0.99. More general models obtained from the combined saturate and aromatics training data sets had lower accuracies. Therefore, separate models developed for either saturates or aromatics alone are recommended for use. The leave-one-out cross validation correlation coefficients were only slightly smaller for all the models reported here, which points to their good quality. Comparisons of model predictions with external test data sets consistently indicated good predictive abilities of all the models. In general, all the models performed well with average prediction errors of less than 1% for both saturates and aromatics and they were dramatically better than the group contributions models examined here.

Densities of hydrocarbons were found to be related to their refractive indices, while no significant correlations were observed between *BP* and *SG*, or *BP* and *RI*. The selected descriptors reflected the characteristics of the hydrocarbons in the data set. Generally, the molecular constitution, topology, geometry, as well as the polarity were the major factors affecting physical properties like *BP*, *SG*, and *RI*. The ring structures and molecular weight in various forms (e.g. Number of Rings, $MW^{1/3}$) were the key constitutional descriptors. Topological indices (i.e., Balaban, Wiener, and Randic) and Information Content indices were important descriptors in predicting the *BP* of saturates. The best geometrical descriptors in the above models are YZ Shadow, a reduced form of the gravitational index ($G^{1/2}$ or $G^{1/3}$), and the Moment of Inertia C. The polar intermolecular interactions in all the models were captured by: CPSA descriptors, Total Point-Charge Component of the Molecular Dipole, and the Maximum (Minimum) Partial Charge of the H Atom. In the *SG* and *RI* models for saturates, and the *RI* model for both saturates and aromatics together, four out of the seven variables were these partial charge descriptors. However, they were less important in the *BP* model for saturates, where only one of them was used (the Polarity Parameter).

2.6 References

API technical data Book, American Petroleum Institute, Washington, D.C, 1995.

Balaban, A. T. (1982). Highly Discriminating Distance-Based Topological Index, *Chem. Phys. Lett.*, 89, 399-404.

Beilstein CrossFire Database, Beilstein, 2000.

Bjørseth, A. Handbook of Polycyclic Aromatic Hydrocarbons, Marcel Dekker Inc., New York, 1983.

Christensen, G.; Apelian, M. R.; Hickey, K. J.; and Jaffe, S. B. (1999). Future directions in modeling the FCC process: An emphasis on product quality, *Chem. Eng. Sci.*, 54, 2753-2764.

CODESSA Reference Manual, SemiChem, KS, USA, 1997.

CRC Handbook of Chemistry and Physics, edited by Lide, D. R., 81st edition, CRC Press, NY, 2000-2001.

Egolf, L. M; Wessel, M. D.; and Jurs, P. C. (1994). Prediction of boiling points and critical temperatures of industrially important organic compounds from molecular structures, *J. Chem. Inf. Comput. Sci.*, 34, 947-956.

Elbro, E. S.; Fredenslund, A.; and Rasmussen, P. (1991). Group contribution method for the prediction of liquid densities as a function of temperature for solvents, oligomers, and polymers, *Ind. Eng. Chem. Res.*, 30, 2576-2582.

Ferreira, M. M. C. (2001). Polycyclic aromatic hydrocarbons: a QSPR study, *Chemosphere*, 44, 125-146.

Joback, K. G. A unified approach to physical property estimation using multivariate statistical techniques, M.S. Dissertation, The Massachusetts Institute of Technology, Cambridge, MA, 1984.

Karcher, W. Spectral Atlas of Polycyclic Aromatic Compounds, vol. 2, Kluwer Academic Publishers, Dordrecht, MA, 1988.

Katritzky, A. R.; and Gordeeva, E. V. (1993). Traditional topological indices *vs* electronic, geometrical, and combined molecular descriptors in QSAR/QSPR research, *J. Chem. Inf. Comput. Sci.*, 33, 835-857.

Katritzky, A. R.; Lobanov, V.; and Karelson, M. CODESSA (Comprehensive Descriptors for Structural and Statistical Analysis), SemiChem and University of Florida, Gainesville, FL, 1997.

Katritzky, A. R.; Mu, L.; Lobannov, V. S.; and Karelson, M. (1996). Correlation of boiling points with molecular struture. 1. A training set of 298 diverse organics and a test set of 9 simple inoraganics, *J. Phys. Chem.*, 100, 10400-10407.

Katritzky, A. R.; Perumal, S.; Petrukhin, R.; and Kleinpeter, E. (2001). CODESSA-based theoretical QSPR model for hydantoin HPLC-RT Lipophilicities, *J. Chem. Inf. Comput. Sci.*, 41, 569-574.

Kier, L. B. (1986). Indexes of molecular shape from chemical graphs, *Acta Pharm. Jugosl.*, 36, 171-188.

Kirpichenok, M. A.; and Zefirov, N. S. (1987). *Zh. Org. Khim.*, 23, 4.

Martens, H. A.; and Dardenne, P. (1998). Validation and verification of regression in small data sets, *Chemometrics and Intelligent Laboratory Systems*, 44, 99-121.

MOPAC2002 User's Manual, Fujitsu Limited, OR, USA, 2000.

Murugan, R.; Grendze, M. P.; Toomey, J. E. Jr.; Katritzky, A. R.; Karelson, M.; Labanov, V.; and Rachwal, P. (1994). Predicting physical properties from molecular structure," *CHEMTECH*, June, 17-23.

Neter, J.; Kutner, M. H.; Nachtsheim, C. J.; and Wasserman, W. Applied Linear Statistical Models, 4th ed., Irwin, Chicago, 1996.

Neurock, M. N.; Nigam, A.; Trauth, D.; and Klein, M. T. (1994). Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms, *Chem. Eng. Sci.*, 49, 4153-4177.

Rackett, H. G. (1970). Equation of state for saturated liquids, *J. Chem. Eng. Data*, 15, 514.

Randic, M. (1975). On characterization of molecular branching, *J. Am. Chem. Soc.*, 97, 6609-6614.

Retzekas, E.; Voutsas, E.; Magoulas, K.; and Tassios, D. (2002). Prediction of physical properties of hydrocarbons, petroleum, and coal liquid fractions, *Ind. Eng. Chem. Res.*, 41,1695-1702.

Riazi, M.R.; and Roomi, Y. (2001). Use of Refractive Index in Estimating Thermo-physical Properties of Hydrocarbon Mixtures, *Ind. Eng. Chem. Res.*, 40, 1975-1984.

Rohrbaugh, R. H.; and Jurs, P. C. (1987). Descriptions of molecular shape applied to studies of structure/activity and structure/property relationship, *Anal. Chim. Acta*, 199, 99-109.

Rossini, F. D. Selected Value of Physical and Thermodynamic Properties, American Petroleum Institute, Carnegie Press, 1953.

Shannon, C. E. (1948). A mathematical theory of communication, *Bell. Syst. Technol. J.*, 27, 379-423.

Smittenberg, J.; and Mulder, D. (1948). Relations between refraction, density, and structure of series of homologous hydrocarbons. I. Empirical formula for refraction and density at 20°C of n-alkanes and n-alpha-alkanes, *Rec. Trav. Chim*, 67, 813-838.

Stanton, D. T.; and Jurs, P. C. (1990). Development and use of charged partial surface area structural descriptors in computer-assisted quantitative structure-property relationship studies, *Anal. Chem.*, 62, 2323-2329.

Stein, S. E.; and Brown, R. L. (1994). Estimation of normal boiling point from Group Contributions, *J. Chem. Inf. Comput. Sci.*, 34, 581-587.

Stuper, A. J.; Brugger, W. E.; and Jurs, P. C. Computer-Assisted Studies of Chemical Structure and Biological Function, Wiley-Interscience, New York, 1979.

Van Nes, K.; and van Westen, H. A. Aspects of the constitution of mineral oils, Elsevier Publishing Co., New York, 1951.

Wessel, M. D.; and Jurs, P. C. (1995). Prediction of normal boiling points of hydrocarbons from molecular structure, *J. Chem. Inf. Comput. Sci.*, 35, 68-76.

Wiener, H. (1947). Structural determination of paraffin boiling points, *J. Am. Chem. Soc.*, 69, 17-20.

Chapter 3

Hydrocarbon Isomer Distribution in Petroleum Mixtures

3.1. Introduction

Characterization of petroleum fractions is critically important to advanced kinetics modeling of various conversions processes in petroleum refining. Recent attempts to model the hydrocracking and catalytic cracking processes require molecular representation of the feedstock (Souverijns et al., 1998; Mizan and Klein, 1999). This trend in process modeling is driven by the desire to predict, in a fundamental way, not only the yields of individual product fractions but also their detailed properties. On the other hand, positive identification and quantification of large numbers of isomers is beyond the capabilities of today's analytical techniques. The isomeric lump frequently sets the limit for molecular characterization (Briker et al., 2001). Even if it were possible to know the exact composition of the fraction, computational limitations make it impossible to use this amount of information. Therefore, usually there is a practical limit as to how much a process modeler may know about feedstock composition. For example, structural-oriented lumping (Quann and Jaffe, 1992) and single-event kinetics (Vynckier and Froment, 1991) models rely on characterization of feedstock in terms of several molecular classes (homologous series) distributed by carbon number. However, a strong assumption in lumping isomers (by type and carbon number) is that the physical and chemical properties of those isomers are identical. This assumption is not true for most of hydrocarbons.

Many thermo-physical properties of various isomers are widely spread. For example, the difference in normal boiling points (NBP) among terphenyls (o-, m-, p-) is 50K. The

maximum difference of normal freezing points among octane isomers is 227K, with a maximum of 374K for 2,2,3,3-tetramethyl butane and a minimum of 147K for 2,3-dimethyl hexane (API Technical Data Book, 1992). If isomeric lumps are considered instead of individual molecules, it is not possible to estimate reliably bulk properties for an arbitrary stream. However, if the molecular makeup of a refinery stream is known at the molecular level, an efficient property estimation model could be used to estimate its properties. One possible way to achieve this is to use a quantitative-structure-property-relationship (QSPR) model to estimate a particular property of each individual hydrocarbon in the stream (Ha et al., 2005a) and then estimate this property for the whole stream using appropriate mixing rules. The chemical activities and reaction paths of hydrocarbons are also dependent on isomer distribution. For example, different isomers produce different carbonium or carbenium ions during catalytic cracking. As a result, they go through different elementary reaction paths. In a catalytic cracking study of three C_6 isoparaffins (2-methyl pentane, 3-methyl pentane, and 2,3-dimethyl butane), Wojciechowski (1998) found that these three C_6 isomers followed quite different reaction paths in the initiation, propagation, and β-cracking. As a result, their corresponding products significantly differed in terms of the kinetic chain length (3.38, 3.12, and 27.03, respectively), paraffin/olefin ratio (3.38, 1.21, and 10.75 respectively), and volume expansion (1.30, 1.83, and 1.09 respectively). Therefore, the distribution of isomers is important in the estimation of bulk physical properties, as well as in the detailed kinetic study of complex mixtures.

The capabilities of analytical techniques rapidly decrease with boiling range. Composition of refinery streams in the naphtha boiling range can be measured at the

molecular level using the DHA (Detailed Hydrocarbon Analysis) method. However, mass spectrometry, probably the most capable method for distillate characterization, is incapable of distinguishing various isomers. Hence, an isomeric lump is the practical limit for the compositional detail available from the analytical laboratory. Although advances in characterization of petroleum fractions benefit from the development of new more advanced analytical techniques, this may not be the only way to deliver the detail necessary for reliable modeling of product quality.

Finding isomer distribution within an isomeric lump has been considered an intractable problem (Kuo, 1991). These distributions reflect the reaction conditions during the crude maturing processes. Consequently, the abundances of individual isomers in the isomeric lump would be expected to depend on the kinetics of the reactions they undergo and their thermodynamic stabilities. Frequently, the thermodynamic equilibrium among isomers has been assumed in isomeric lumping for kinetics modeling (Krambeck, 1991), and equilibrium distribution has been analytically solved for ideal solutions (Smith and Missen, 1982). However, the actual distribution of isomers differs from the equilibrium distribution in most cases, especially for saturates (Tissot and Welte, 1984). Although it is infeasible to quantify each individual isomer in a large isomeric lump (e.g. $>C_{10}$), only a relatively small fraction of the set of all possible molecules is actually present in various petroleum fractions in quantities that affect their processability and, ultimately, quality (Tissot and Welte, 1984).

This work proposes a deterministic way of finding isomer distribution in an isomeric lump independent of the limitations of analytical methods. We found that the distribution of isomers in the isomeric lump could be calculated by minimizing the Gibbs free energy

of the lump subject to a constraint in addition to the stoichiometric one; the independently measured boiling point distribution within this lump. By default, this boiling point distribution is measured with decreasing degree of accuracy for isomer systems of increasing carbon number. The approach was applied to estimate the hexane and heptane isomer distributions and the results were compared to the distributions in light petroleum fractions published in the open literature. The validity of this approach and the uniqueness of the solution to the associated mathematical problem were examined. The proposed approach provides a novel and efficient method for determining the distribution of isomers, which reflects the thermodynamic stability aspects of molecular composition.

3.2 Isomer distribution within an isomeric lump

Isomer distribution, so far, has been studied at thermodynamic equilibrium (Smith and Missen, 1982; Alberty, 1991) To simplify the problem, the isomeric lump is frequently assumed to be a closed ideal system. This approach is also taken here. In such a system, the Gibbs free energy of an ideal solution of N isomers can be expressed as:

$$\Delta G_n^o = \sum_{i=1}^{N} x_i \Delta G_i^0 + RT \sum_{i=1}^{N} x_i \ln x_i \qquad (3\text{-}1)$$

subject to the stoichiometric constraint $\sum x_i = 1$. The equilibrium composition of the isomeric lump can be obtained by minimizing ΔG_n^o with respect to the system composition (x_i) to give (Alberty, 1991)

$$x_i = \exp[(\Delta G_n^o - \Delta G_i^o)/RT] \qquad (3\text{-}2)$$

where ΔG_n^o is defined as

$$\Delta G_n^o = -RT \ln\left[\sum_{i=1}^{N} \exp(-\Delta G_i^o / RT)\right] \qquad (3\text{-}3)$$

Using this approach and based on limited data available in the open literature, the measured isomeric distribution in virgin crude oils is not consistent with thermodynamic equilibrium. Martin et al. (1963) measured distributions of small alkane isomers in naphtha from 18 crude oils. The averaged distribution (they found little variation in heptane isomer distributions among these crudes) is compared with the calculated equilibrium distributions in Table 3.1 for 298K and 400K, with respect to the gas and liquid phases. (Tissot and Welte, 1984) Clearly, the match between the measured and equilibrium distributions is inadequate. The same was found to be true for hexane isomer distributions (Martin et al., 1963).

Table 3.1 Abundances of heptane isomers in virgin crude oils (Tissot and Welte, 1984)

Isomers/Distribution	Abundance, wt%		Isomerization equilibrium, wt%			
	Ave. 18 oils[#]	HM[*]	298K(g)	298K(l)	400K(g)	400K(l)
n-heptane	55.5	52.0	1.25	2.25	4.3	8.55
2-methyl hexane	13.8	16.0	9	11.2	15.4	15.65
3-methyl hexane	19.2	22.4	5.1	6.85	11.3	12.15
3-ethyl pentane	2.6	2.4	0.45	0.6	1.3	1.45
2,2-dimethyl pentane	0.6	0.4	32	24.9	16.7	13.2
2,3-dimethyl pentane	6.1	4.8	25	30	28.8	29.5
2,4-dimethyl pentane	1.7	2.0	9.9	8.15	8.4	6.8
3,3-dimethyl pentane	0.4	-	11.4	11.3	10.4	9.8
2,2,3-trimethyl butane	0.1	-	5.9	4.75	3.4	2.9

*Hassi-Messaoud crudes (monophasic sample); [#] Martin et al. (1963),

A solution to this problem is proposed below. We assume that all the existing isomers are in a state that can be estimated by considering the classical equilibrium problem, subject to the stoichiometric constraint, with an additional constraint of partial information about the system composition. One way to obtain this partial information would be to measure the boiling point distribution of the mass in the isomeric lump by an

appropriate GC technique. Note that detailed (rather than partial) information about boiling point distribution would be equivalent to knowing the system composition in detail. This partial information could be the average boiling point of the lump, if the lump consists of a relatively small number of isomers with widely spread (relatively to the accuracy of measurement) boiling points. The average boiling point measured with finite accuracy may not provide a sufficient amount of information for a lump with a relatively narrow boiling point spread or consisting of a large number of isomers. In those cases, as much information as possible about the boiling point distribution should be provided.

The concentration of structural isomeric lumps (SILs), defined as hydrocarbon species of the same carbon number within a hydrocarbon homologous series, can be quantified using relatively low-cost advanced analytical techniques such as GC-FIMS. With the help of appropriate GC retention time calibration, each SIL can be assigned a boiling point distribution. The use of n-paraffin standards to link the boiling points with retention time (basis of the ASTM D2887 simulated distillation method) is assumed to be sufficiently accurate in this work but, in principle, other more sophisticated methods could be used for retention calibration for individual hydrocarbon groups (e.g. retention calibration for each individual hydrocarbon type). The boiling point differences between individual isomers are reflected in differences between their retention times. However, it should be noted here that when the number of isomers is large, usually it is not possible to resolve their corresponding individual peaks by standard chromatography. Therefore, the boiling point distribution of a SIL cannot be measured in the detail required for determination of its composition and other sources of information are required to achieve it (e.g. methodology proposed in this paper).

If detailed molecular composition of the SIL is known, for example, through simulation, its boiling point distribution or an average boiling point (BP_{lump}) in less complex cases, can be obtained from the boiling points of individual isomers and their concentrations (in case of BP_{lump}, an appropriate mixing rule is used). Therefore, operationally, partial information about the SIL composition can be used as a constraint to estimate the isomer distribution and, as proposed here, the problem becomes one of constrained minimization of ΔG_n^o defined by Equation 3-1

$$Min.\left[\Delta G_n^0\right] = Min.\left[\sum_{i=1}^{N} x_i \Delta G_i^0 + RT\sum_{i=1}^{N} x_i \ln x_i\right]$$
$$Subject\ to\ \begin{cases} \sum x_i = 1 \\ \sum x_i BP_i = BP_{lump} \end{cases} \qquad (3\text{-}4)$$

The uniqueness of the solution for this problem in a general case is shown in Appendix B. The calculations discussed below, conducted using Powell's method (1989) (a modified Newton and Raphson method), yielded the composition vector (X) of dimension N, predicting the hexane and heptane isomer distributions presented in the following section.

3.3. Simulated isomer distributions and discussion

Minimization of Equation 3-4 subject to the stoichiometric and BP_{lump} constraints was applied to simulate the isomeric distribution of the hexane and heptane isomers. Table 3.2 lists the densities, boiling points, and free energies of formation in the gas and liquid phases for each isomer used in the calculations. The densities, boiling points, and free energies of formation in the gas phase are reported in the API Technical Data Book

(1992). The free energies of formation in the liquid phase were calculated from the standard free energies of formation in the gas phase, reported heat of vaporization, heat capacities of gas and liquid phases, and the entropies of gas and liquid phases at standard state. Since the original GC data for the reported hexane and heptane isomer distributions were not available, the BP_{lump} was estimated directly from the actual concentration distribution instead. Again, normally, BP_{lump} can be calculated from the GC-MS data. Distributions of hexane and heptane isomers in the gas and liquid phases were calculated using the free energy of formations at standard state in the gas and liquid, respectively.

Table 3.2 Properties of hexane and heptane isomers and average predicted results

Isomers\Properties	$\Delta G^0_{g,\,298K}$ kcal/mol	$\Delta G^0_{l,\,298K}$ kcal/mol	BP, °C	$\rho_{15°C}$, g/ml	predicted wt% 298K(g)	predicted wt% 298K(l)	abund*. wt% in 18 crudes
n-hexane	-0.016	-1.03	68.73	0.6651	53.778	56.821	52.787
2-methyl pentane	-1.275	-1.97	60.26	0.6577	19.650	17.559	24.916
3-methyl pentane	-0.512	-1.34	63.27	0.6693	20.729	17.064	18.581
2,2-dimethyl butane	-2.089	-2.90	49.73	0.6539	1.803	2.838	0.591
2,3-dimethyl butane	-0.7464	-1.69	57.98	0.6662	4.040	5.719	3.125
n-heptane	1.9515	0.3564	98.43	0.690	56.157	56.590	55.5
2-methyl hexane	0.8294	-0.5631	90.05	0.682	15.925	15.510	13.8
3-methyl hexane	1.2247	-0.2176	91.85	0.692	16.099	15.961	19.2
3-ethyl pentane	2.7199	1.3215	93.47	0.704	2.379	2.064	2.6
2,2-dimethyl pentane	0.1315	-1.2331	79.19	0.682	0.870	1.204	0.6
2,3-dimethyl pentane	1.3664	-0.072	89.78	0.699	5.822	6.186	6.1
2,4-dimethyl pentane	0.8142	-0.341	80.49	0.676	0.449	0.416	1.7
3,3-dimethyl pentane	1.1735	-0.0875	86.06	0.696	1.988	1.796	0.4
2,2,3-trimethyl butane	1.1186	-0.0139	80.88	0.695	0.311	0.273	0.1

*Averaged abundance in reported 18 crude oils (Martin et al., 1963)

The normalized simulated distributions of hexane and heptane isomers are compared below to the distributions measured by Martin et al. (1963) for 18 crude oils. These crude oils spanned a wide range of geological ages and represented compositional extremes

with API gravities ranging from 18 to 45. Eleven were found in Paleozoic rocks, two in Mesozoic, and five in Cenozoic. Some important characteristics of the 18 crude oils are given in Table 3.3. Further details can be found in the original publication by Martin et al. (1963). The hexane and heptane isomers had been quantified in the naphtha fractions boiling up to 111°C by GC using a capillary column. The hexane and heptane isomers were quantified in vol% of this naphtha fraction. The total amounts of hexane or heptane isomers were between 2 and 4% in most of the crude oils. Densities at 15°C were used to convert the vol% to wt% used in this work. The estimated uncertainties in the original results were less than 6% of the amount reported, or one in the last digit, whichever was larger (Martin et al, 1963).

Table 3.3 General descriptions of 18 crude oils (Martin et al., 1963)

Field	State/Country	°API	Era	Temp. °C	Vol% up to 111°C
Alida	Saskatchewan	38.1	Paleozoic	38	19.21
Bever Lodge	N. Dakota	40.7	Paleozoic	NA	18.43
Darius	Iran	29.0	Mesozoic	118	9.18
Eola Mclish	Oklahoma	36.6	Paleozoic	78	15.69
Eola oil creek	Oklahoma	44.6	Paleozoic	75	22.83
Hendricks	Texas	33.1	Paleozoic	30	14.07
Kawkawlin	Michigan	35.0	Paleozoic	32	10.47
Lee Harrison	Texas	25.3	Paleozoic	NA	14.30
North Smyer	Texas	43.2	Paleozoic	66	27.33
Pembina	Alberta	41.6	Mesozoic	52	21.13
Ponca city	Oklahoma	42.0	Paleozoic	57	17.09
Redwater	Alberta	35.3	Paleozoic	49	17.54
South Houston	Texas	24.0	Cenozoic	58	3.54
Swanson River	Alaska	31.3	Cenozoic	66	12.11
Teas	Texas	40.0	Paleozoic	64	23.46
Uinta Basin	Utah	30.6	Cenozoic	NA	4.50
Wafra	Kuwait	18.3	Cenozoic	NA	4.58
Wilmington	California	19.3	Cenozoic	54	4.44

Table 3.2 also compares the normalized simulated isomeric distributions with the average distribution measured by Martin et al. (1963). Clearly, the agreement is quite close and much better than the agreement between the equilibrium and measured

distributions presented in Table 3.1. The hexane and heptane isomeric distributions were estimated at temperatures of the individual reservoirs. For those whose reservoir temperatures were not available (4 samples), 60°C (the averaged temperature for the remaining 14 crude oils) was used. The free energies of formations for individual hexane and heptane isomers at reservoir temperatures were estimated from the heats of formations and entropies at 300K and 400K (Stull et al., 1969). The simulated hexane and heptane isomer distributions are compared with the reported distributions of 18 virgin crude oils. Good agreements were observed between the predicted heptane-isomer distributions and the reported ones for 15 samples. To be concise, only one of them (Alida) was shown in Figure 3.1. The rest of good matches are illustrated in Appendix B, Figures B1 to B14. The predictive deviations for both hexane and heptane distributions are tabulated in Table 3.4. Similarly good agreements between the simulated and measured distributions were also found for 5 hexane isomers (see Table 3.4). Since the distribution of possible isomers within a SIL is only related to the relative values of their free energies of formations, the distribution pattern of the hexane and heptane isomers in the gas phase were expected to be similar to those in the liquid phase. Indeed, because the relative values of free energy of formations among hexane and heptane isomers in the gas phase were found to be very close to those in the liquid phase, similar distributions in gas and liquid phase were obtained for all 18 crudes (see Figures 3.1, and Figures B1 to B14 for heptanes). Consequently, only the simulated isomeric distributions in the gas phase were used to estimate the prediction errors.

Table 3.4 The prediction deviations (calculated – measured) for normalized distribution of hexane/heptane isomers (wt%)

Crudes /Isomers	Hex.	2M-Pen.	3M-Pen.	22M-But.	23M-But.	Ave abs isomers	Hep.	2M-Hex.	3M-Hex.	3E-Pen.	22M-Pen.	23M-Pen.	24M-Pen.	33M-Pen.	223M-But.	Ave abs isomers
Alida	2.378	-5.898	0.337	1.683	1.500	2.359	1.707	0.987	-5.266	0.870	0.453	-0.234	-0.832	1.995	0.320	1.407
Bever Lodge	0.966	-5.434	2.030	1.058	1.381	2.174	-0.008	0.858	-3.285	1.057	0.038	0.711	-1.184	1.601	0.213	0.995
Darius	2.063	-3.624	-0.883	1.118	1.326	1.803	0.768	-1.339	-4.043	2.057	0.376	0.949	-0.866	1.761	0.339	1.389
Eola Mclish	2.680	-7.198	0.535	1.843	2.140	2.879	2.150	0.354	-5.535	0.027	0.570	1.109	-0.913	1.958	0.281	1.433
Eola oil creek	-0.001	-5.083	3.170	0.6[illegible]	1.288	2.034	-0.999	-0.443	-0.976	1.173	-0.111	1.172	-1.189	1.244	0.128	0.826
Hendricks	8.748	-4.326	-9.515	4.095	0.999	5.537	8.607	1.862	13.624	-1.31	2.730	-1.786	-0.902	3.473	0.951	3.916
Kawkawlin	-1.839	-2.566	4.689	-0.097	-0.187	1.875	-1.081	-0.671	1.061	0.833	-0.110	0.225	-0.366	0.129	-0.020	0.500
Lee Harrison	6.388	-4.981	-4.515	4.048	-0.941	4.174	4.115	0.803	-5.123	-0.14	1.251	-3.767	-0.342	2.684	0.520	2.083
North Smyer	7.502	-9.587	-5.119	3.838	3.366	5.882	5.275	3.185	-8.860	-0.67	1.373	-3.263	-0.623	2.892	0.691	2.981
Pembina	2.654	-5.698	0.036	1.904	1.104	2.279	-0.049	-0.081	-3.485	1.075	0.323	1.508	-2.162	2.430	0.441	1.284
Ponca city	1.333	-5.815	1.637	1.178	1.666	2.326	0.242	0.967	-2.721	0.904	0.123	-0.227	-0.876	1.399	0.188	0.850
Redwater	4.752	-7.538	-2.290	2.636	2.440	3.931	2.557	2.243	-5.819	-0.85	0.627	-0.159	-1.104	2.160	0.348	1.763
South Houston	-3.898	11.257	3.666	0.389	-11.413	6.125	2.214	10.378	0.329	-3.73	4.921	10.270	-4.577	1.417	-0.679	4.280
Swanson River	3.067	-7.122	0.526	2.362	1.168	2.849	-0.484	0.795	-1.342	0.203	0.737	0.156	-2.758	2.350	0.343	1.019
Teas	2.963	-6.502	-0.094	2.002	1.631	2.638	1.503	0.524	-4.702	0.762	0.475	0.005	-1.080	2.152	0.361	1.285
Uinta Basin	3.697	-9.571	0.763	2.658	2.453	3.828	0.215	1.018	-1.908	0.41	0.070	-0.613	-0.900	1.537	0.176	0.760
Wafra	18.426	1.834	-25.881	7.908	-2.288	11.267	15.557	3.864	19.855	-3.23	5.664	-7.639	-1.660	5.195	2.101	7.196
Wilmington	6.996	-4.391	-7.097	3.347	1.145	4.595	5.587	7.378	-5.662	-4.07	1.618	-7.414	-1.179	3.068	0.672	4.072
ARD* in 15 crude oils	0.081	0.223	0.165	4.550	0.547	N/A	0.051	0.076	0.204	0.434	1.156	0.151	0.630	5.236	3.683	N/A

*Ave. Rel. Dev. = $\sum$[|predicted-measured|/measured]/15, South Houston, Wafra, Wilmington are excluded

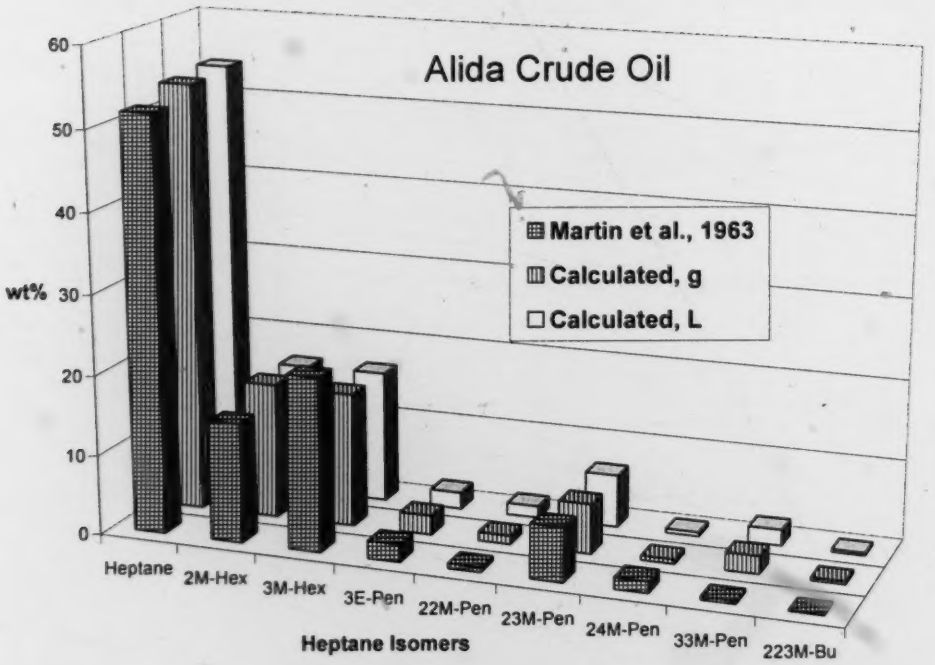


Figure 3.1 Prediction of heptane isomer distribution in Alida crude oil

As shown in Table 3.4, the average absolute errors over all hexane and heptane isomers are less than 3% and 1.5% for most of the reported crudes. However, substantial predictive errors are found for the younger crudes, especially for South Houston, Wafra, and Wilmington (see Figure 3.2 through 3.4 for heptane-isomer distributions), which were formed during the tertiary Cenozoic Era. The maximum average predictive errors are 11% and 7% for hexane and heptane isomers, respectively, in Wafra crude (see Table 3.4). The potential significance of the current approach lies on the fact that the approach was able to predict the key isomers (n-hexane, 2-methyl-pentane, 3-methyl-pentane for hexane isomers; n-heptane, 2-methyl-hexane, 3-methyl-hexane, and 2,3-dimethyl-pentane for heptane isomers) with good confidence. As shown in Table 4, the relative predictive errors were 8.1%, 22.3%, and 16.5% for n-hexane, 2-methyl-pentane, and 3-methyl-

pentane, respectively. Those for n-heptane, 2-methyl-hexane, 3-methyl-hexane, and 2,3-dimethyl-pentane were 5.1%, 7.6%, 20.4%, and 15.1%, respectively. The 3 key isomers made up more than 92% of total hexane isomers except for two younger crudes (South Houston and Wafra), while the 4 key isomers comprised of more than 90% of total heptane isomers, except for those three younger crudes (South Houston, Wafra, and Wilmington). Considering the overall experimental uncertainties (6% or one in the last digit) and the wide distribution range of the key isomers in those crudes, the predictions for key isomers are satisfactory.

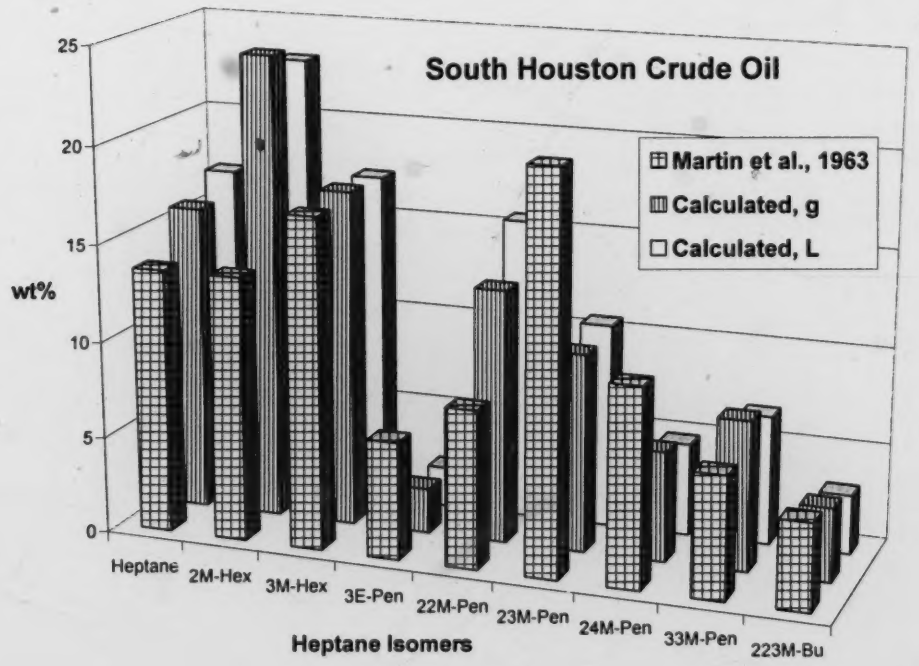


Figure 3.2 Prediction of heptane isomer distribution in South Houston crude oil

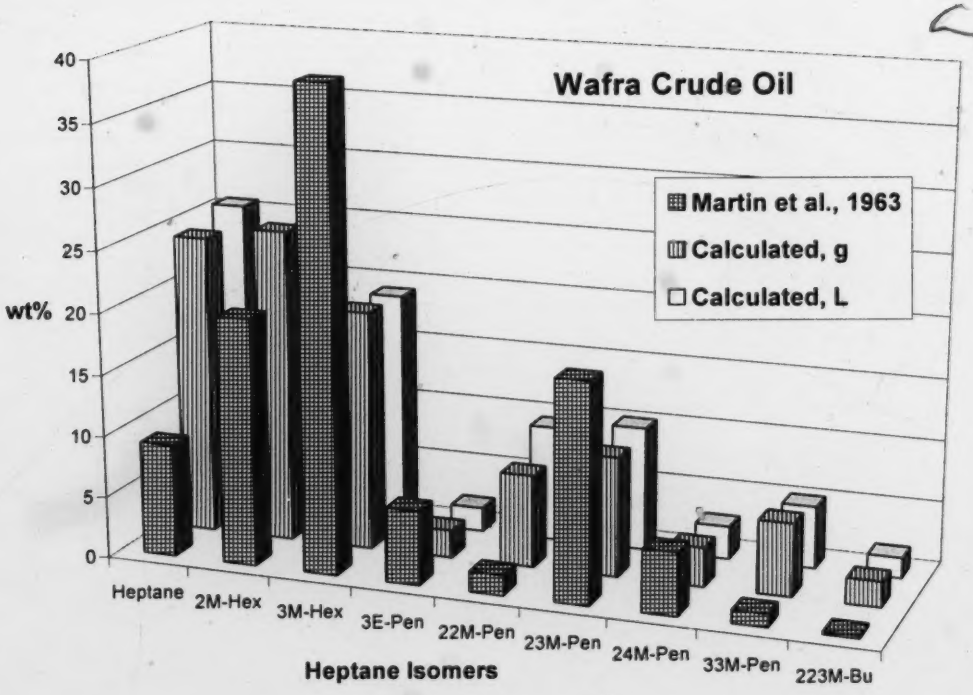


Figure 3.3 Prediction of heptane isomer distribution in Wafra crude oil

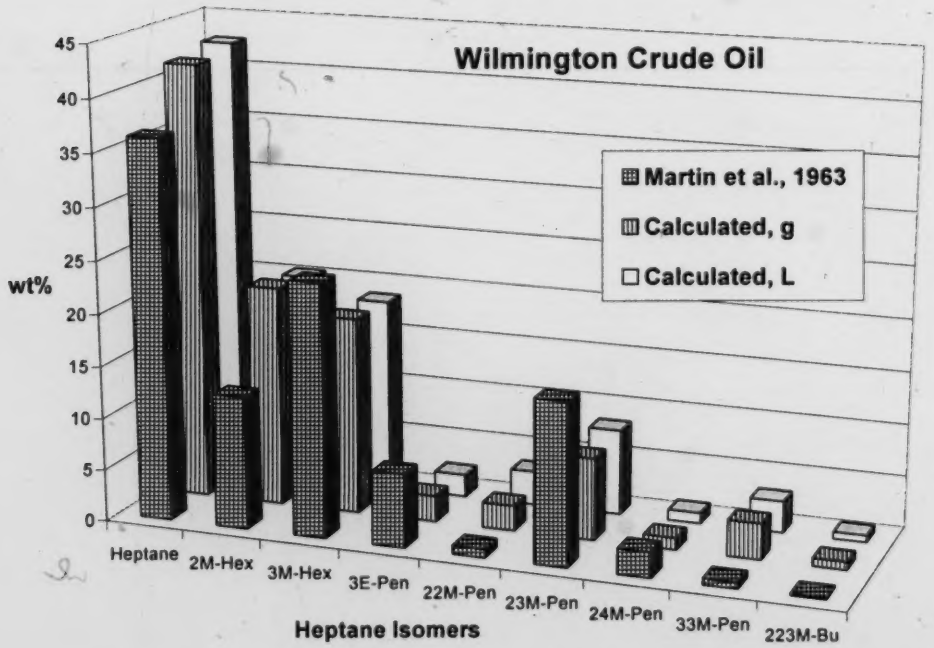


Figure 3.4 Prediction of heptane isomer distribution in Wilmington crude oil

According to Martin et al. (1963), the high proportion of branched paraffins in the three outlier crudes suggested possible contribution of catalytic cracking in their early formation period. These authors suggest that thermal cracking was the dominant process that contributed to the isomeric distributions in naphtha of the mature crudes. In addition to this dissimilarity in the maturation processes, the total amounts of hexane/heptane isomers in these three crudes are relatively small (0.335/0.381%, 0.885/0.847%, and 0.529/0.495% in South Houston, Wafra, and Wilmington crudes, respectively) compared to those of the rest (>2.0%). This likely resulted in larger measurement errors. Consistent with their relative immaturity, the three crude oils were also the three highest boiling mixtures among the 18. Although our proposed approach resulted in larger prediction errors for the 3 crudes than for the remaining 15 (e.g., average absolute errors of 4.28%, 7.2% and 4.07% in normalized heptane isomer distributions), our results are much better than the thermodynamic equilibrium distributions calculated from Equations 2 and 3. Those errors are as high as 7.02%, 11.86% and 13.22% for these three crudes, respectively. The good agreements between the simulated and experimental distributions for the mature crude oils suggest that the proposed method may also be applicable to the predictions of isomer distributions in thermally cracked materials such as coker gas oils.

The purpose of this chapter was not to devise a method for distributing hexane and heptane isomers but to look for a method that would potentially be generally applicable to any SIL, and particularly for heavier SILs. For heavier materials, isolation and identification of individual molecules are infeasible. However, the boiling point distribution can be obtained experimentally for the individual SILs during GC-MS measurements regardless of their boiling range. Although the number of possible

isomeric permutations is very high for larger hydrocarbons, the actual distribution of isomers is not as diverse as might be expected (Hood et al., 1959). As shown by Martin et al. (1963), the distributions of hexane/heptane isomers were dominated by 3 or 4 key isomers that accounted for 92/90% of total SIL. The distributions of these key isomers can be estimated with high accuracy using current approach. It is possible to use a limited number of major isomers to represent a SIL in molecular characterizations without losing the intrinsic chemistry detail (Liguras and Allen, 1989). The selection of major isomers can be based on the analyzed structural occurrences and thermodynamic stabilities of individual molecules. The free energy of formations for larger hydrocarbons can be estimated by means of computational chemistry software packages, such as MOPAC, whenever such experimental data are not available. Consequently, at least operationally, it is possible to simulate the isomer distribution of larger hydrocarbons using the proposed approach. The validity of this approach in the generation of molecular composition of higher boiling fractions is explored elsewhere (Ha et al., 2005b).

3.4 Conclusions

A computational augmentation for GC-MS techniques has been proposed to determine the composition of petroleum samples in detail, impossible to achieve through GC-MS or any other analytical techniques alone. The proposed methodology uses a partial knowledge of the composition of a structural isomeric lump introduced in the form of a constraint derived from its boiling point distribution in the calculation of the thermodynamic equilibrium among the isomers involved. The resulting molecular distribution closely resembled experimental distributions when this approach was applied

to predicting the isomeric distribution among hexane and heptane isomers reported for 18 geologically different crude oils. Excellent agreement was found for 15 crudes, while a worse, but still favourable agreement was found for the remaining three younger (Cenozoic) crudes. Although, in view of very scarce experimental data, this approach could only be tested using data on isomer distributions in the naphtha boiling range, it has potential applications in predicting the isomer distributions for heavy fractions (Ha et al., 2005b). The BP distribution constraint used in the Gibbs free energy minimization captures partial knowledge about the composition of the SIL under consideration, regardless of the boiling point range. It was demonstrated that, for simpler SILs with a boiling point spread that is wide enough (compared with the accuracy of the boiling point distribution measurement), as little compositional information as the average boiling point of the SIL carries sufficient information to help find the isomeric distribution with excellent accuracy. For SILs that involve more isomers with less boiling point spread, the boiling point distribution (rather than the average boiling point) may be required to achieve the same accuracy. It has been demonstrated here that the composition of petroleum mixtures, going well beyond the capabilities of the analytical methods available today, can be determined with good accuracy through a computer simulation.

3.5 References

Alberty, R. A. Chemical equilibrium in complex organic systems with carious choices of independent variables. *Chemical Reactions in Complex Mixtures*, edited by Sapre, A. V. and Krambeck, F. J. Van Nostrand Reinhold, New York, 1991.

API Technical Data Book – Petroleum Refining, 5th edition, American Petroleum Institute, Washington, D. C., May, 1992.

Briker, Y.; Ring, Z.; Iacchelli, A.; McLean, N.; Rahimi, P. M.; and Fairbridge, C. (2001). Diesel fuel analysis by GC-FIMS: aromatics, n-paraffins, and isoparaffins. *Energy & Fuels*, **15**: 23-37.

Ha, Z.; Ring, Z.; Liu, S. (2005a.). Quantitative Structure-property Relationship (QSPR) models for boiling points, specific gravities, and refractive indices of hydrocarbons, *Energy & Fuels*, **19**, 152-163.

Ha, Zhanyao; Liu, Shijie; and Ring, Z. (2005b). Derivation of molecular representations of diesels, **19**(4), in publication.

Hood, A.; Clere, R. J.; and O'Neal, M. J., (1959). The molecular structure of heavy petroleum compounds. *J. Inst. Petrol.* **45**, 168-173.

Krambeck, F. J. An industrial viewpoint on lumping. Kinetic and thermodynamic lumping of multicomponent Mixtures, edited by Astarita, G. and Sandler, S. I., Elsevier Science Publishers B. V., Amsterdam, 1991.

Kuo, J. C. W. Uses and needs of thermodynamics in the oil industry. *Chemical Reactions in Complex Mixtures*, edited by Sapre, A. V. and Krambeck, F. J. Van Nostrand Reinhold, New York, 1991.

Liguras, D. K.; and Allen, D. T., (1989). Structural model for catalytic cracking. 2. Reactions of simulated oil mixtures. *Ind. Eng. Chem. Res.*, **28**, 674-683.

Martin, R. L.; Winters, J. C.; and Williams, J. A., (1963). Composition of crude oils by gas chromatography: geological significance of hydrocarbon distribution. 6th World Pet. Congr., **Sec. V**, 231-260.

Mizan, T. I; Klein, M. T, (1999). Computer-assisted mechanistic modeling of n-hexadecane hydroisomerization over various bifunctional catalysts. *Catalysis Today*, **50**, 159-172.

Neurock, M. N.; Nigam, A.; Trauth, D.; and Klein, M. T. (1994). Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms, *Chem. Eng. Sci.*, **49**, 4153-4177.

Powell, M. J. D. TOLMIN: A fortran package for linearly constrained optimization calculations, DAMTP Report NA2, University of Cambridge, England, 1989.

Quann, R. J; and Jaffe, S. B. (1992). Structure-oriented lumping: Describing the chemistry of complex hydrocarbon mixtures, *Ind. Eng. Chem. Res.*, **31**, 2483-2497.

Smith, W. R.; and Missen, R. W. Chemical reaction equilibrium analysis: theory and algorithm. John Wiley & Sons, New York, 1982.

Souverijns, W.; Martens, J. A.; Froment, G. F.; and Jacobs, P. A., (1998). Hydrocracking of Isoheptadecanes on Pt/H–ZSM-22: An Example of Pore Mouth Catalysis, *Journal of Catalysis*, **174,** 177–184.

Stull, D. R.; Westrum Jr., E. F.; and Sinke, G. C. *The chemical thermodynamics of organic compounds*, John Wiley & Sons, Inc. New York, 1969.

Tissot, B. P.; and Welte, D. H. Petroleum Formation and Occurrence, 2nd edition, Springer-Verlag, Berlin, 1984.

Vynckier, E., Froment, G. F. Kinetic and thermodynamic lumping of multicomponent Mixtures, edited by Astarita, G. and Sandler, S. I., Elsevier Science Publishers B. V., Amsterdam, 1991.

Wojciechowski, B. W. (1998). The reaction mechanism of catalytic cracking: quantifying activity, selectivity, and catalyst decay. *Catal. Rev.-Sci. Eng.*, **40**, 209-328.

Chapter 4

Data Analysis and Reconciliation

4.1 Introduction

As stated in the previous chapters, the ultimate goal of this thesis was to develop a detailed molecular characterization methodology beyond the analytical limitations. To succeed, one must start from the most detailed compositional profile afforded by current analytical techniques, which is the hydrocarbon type distribution by #C measured by HPLC-FIMS (Boduszynski, 1988), GC/LC-MS (Chasey and Aczel, 1991), or GC-FIMS (Briker et al., 2001[a]). A GC-FIMS method for detailed hydrocarbon type characterization of diesel fuels has been developed at the National Centre for Upgrading Technology (NCUT) (Briker et al., 2001[a-b]). The method produces a detailed 19 homologous series by #C distribution matrix (19×17). The hydrocarbon types include n-paraffins, isoparaffins, 3 cycloparaffin subgroups (monocycloparaffin, dicycloparaffin, and polycycloparaffin), 3 monoaromatics (alkylbenzene, benzocycloalkane, and benzodicycloalkane), 3 diaromatics (naphthalene, naphthocycloalkane, and fluorene), 2 triaromatics (phenanthrene and phenanthro-cycloalkane), 2 tetraaromatics (pyrene and chrysene), and 3 aromatic sulfur groups (benzothiophene, dibenzothiophene, and benzonaphtho-thiophene). However, discrepancies have been reported for quantifying the light ends of isoparaffins between GC-FIMS and other standard methods. The difference may be due to the potential fragmentation of the small isoparaffins in FIMS, which are the easiest molecule to be fragmented (Briker et al., 2001[a]). Therefore, for the samples that have a substantial amount of light ends (<200°C), PIONA analysis has been suggested for determining the hydrocarbon type distribution of that light fraction (<200°C), and then

use GC-FIMS to quantify the fractions boiling above 200°C. Integration of PIONA and GC-FIMS results is required for a completed compositional profile of a measured sample. The integrating strategy will be discussed in this chapter.

Bulk properties (e.g. MW, density, CHNS content) and SimDis have been frequently used for verifying molecular characterization methods (Neurock et al., 1994; Sheremata et al., 2004). In this project, physical properties, MW, density, and Refractive Index (RI), are used to validate the developed characterization method. The SimDis results are also used to check the BP distribution of the selected representative molecules. Applying the n-papraffin (RT-BP) standard to the GC-FIMS measurements, a similar FIMS report, hydrocarbon type by BP distribution, can be generated. Summing up all the hydrocarbon types on each BP interval (10°C) results in an equivalent SimDis curve "FIMS-Gen SimDis". Putting the GC-FIMS measurements on the BP scale, one can check the consistency between the GC-FIMS measurements and the SimDis results. This consistency check enables an internal data reconciliation between GC-FIMS and SimDis measurements. It also ensures the accuracy in GC-FIMS results. As a result, the molecular make-up developed from the GC-FIMS results can be compared to the SimDis data, again to validate the molecular characterization approach proposed. These "data reconciliation" procedures are implemented on each sample used for molecular characterization.

4.2 Compositional analyses and bulk property measurements

Five, compositionally different, diesels were characterized to validate our molecular-representation technique. Samples S1 through S3 were obtained from an Edmonton

Refinery: Sample S1 was a heavy naphtha (mainly boiling between 130-230°C) with a high paraffinic content, Sample S2 was a highly paraffinic light distillate (mainly boiling between 200-300°C), and Sample S3 was highly naphthenic. Sample S4 was a hydrotreated diesel derived from Canadian bitumen with a high naphthenic content. Sample S5 is a highly aromatic light distillate from an off-shore crude. These samples were analyzed for MW, density, RI, SimDis, and hydrocarbon class analysis by GC-FIMS. Since all five samples contain a substantial amount of light fractions (<200°C), PIONA analysis was applied to determine the hydrocarbon type profile of the light fractions to avoid the uncertainties in quantifying small isoparaffins in GC-FIMS, whereas GC-FIMS was used for hydrocarbon type quantification of the fractions boiling above 200°C. These bulk properties and main hydrocarbon-type contents are summarized in Table 4.1. All the analytical experiments were conducted at NCUT. Density at 15.6°C was measured by ASTM D4052 using a DMA4 PAAR Densitometer (Annual Book of ASTM Standards, 2001). RI at 25°C was tested by ASTM D1218 using an ABBE Refractometer (Annual Book of ASTM Standards, 2001). MW was determined by Freezing Point Depression (FPD) method using an NCUT internal standard procedure. SimDis was conducted using ASTM D2887 procedure with a HP 6890 GC (Annual Book of ASTM Standards, 2001).

Table 4.1 Summary of the bulk properties and main components for five diesel samples

Sample	Sat*	nP/iP	CycP	Aro	1As	2As	3As	Aro-S	IBP/FBP, °C	RI	Density	MW
S1	76.98	46.56	30.42	23.02	22.24	0.74	0.00	0.04	97/277.3	1.4436	0.7985	150
S2	73.52	33.60	39.92	26.48	20.83	5.58	0.01	0.06	121/344.5	1.4593	0.8291	188
S3	79.51	5.58	73.92	20.49	19.28	1.19	0.00	0.02	118/334.1	1.4584	0.8357	176
S4	74.17	8.48	65.69	25.83	24.41	1.37	0.00	0.05	114.8/338	1.4652	0.8476	176
S5	54.74	18.62	36.12	45.26	33.98	11.18	0.00	0.10	152/299.5	1.4755	0.8537	175

*Sat: saturates, nP: normal paraffins, iP: isoparaffins, CycP: cycloparaffins, Aro: aromatics, 1As: monoaromatics, 2As: diaromatics, 3As: triaromatics, Aro-S: aromatic sulfur

The PIONA method provides for the determination of n-paraffins, iso-paraffins, olefins, naphthenes, and aromatics by carbon number in hydrocarbon streams having final BP of 200°C or less. The test was conducted by an Analytical Control PIONA analyzer — a modified HP5890 GC controlled by Hewlett-Packard chemstation software. The PIONA result for sample 1 is shown in Table 4.2.

Table 4.2 PIONA report for sample 1 (Heavy Naphtha)

#C	Naphthenes	Iso-Paraffins	n-Paraffins	Aromatics	Totals
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00
6	0.10	0.02	0.03	0.01	0.17
7	0.72	0.17	0.28	0.25	1.42
8	1.61	0.74	0.78	1.74	4.88
9	3.95	2.05	2.75	8.09	16.84
10	5.21	8.39	5.54	3.18	22.32
11	2.65	6.20	4.82	0.00	13.67
Totals	14.23	17.57	14.21	13.28	59.29

The detailed hydrocarbon-type analysis by GC-FIMS is a fast method developed by NCUT (Briker et al., 2001[a-b]). Without prior separation of the sample, this method gives a detailed hydrocarbon type distribution from C5 up to C21. The GC-FIMS report of sample 1 is illustrated in Table 4.3 as an example. Sample 1 is heavy naphtha boiling between 97°C and 277°C, corresponding to C6 and C15 respectively. As mentioned above, the report accounts for the fractions of sample 1 boiling above 200°C for better accuracy. Since tetraaromatics and benzonaphthothiophenes were not identified in the selected diesel samples, 16 hydrocarbon classes were reported instead. The PIONA data and GC-FIMS reports for samples 2-5 are tabulated in Appendix C, Tables C1-C8, with the PIONA for the light ends (<200°C) the GC-FIMS for the other tables.

Table 4.3 GC-FIMS by #C distribution report normalized for >200°C fractions of sample 1

HC Type / #C	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Sum
Saturates	**0.00**	**0.00**	**0.00**	**0.00**	**0.12**	**6.47**	**31.39**	**24.63**	**11.97**	**1.47**	**76.07**
Paraffins	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**16.71**	**12.12**	**6.49**	**0.99**	**36.31**
isoparaffins	0.00	0.00	0.00	0.00	0.00	0.00	10.93	8.42	5.58	0.95	25.88
n-Paraffins	0.00	0.00	0.00	0.00	0.00	0.00	5.78	3.70	0.92	0.04	10.43
Cycloparaffins	**0.00**	**0.00**	**0.00**	**0.00**	**0.12**	**6.47**	**14.68**	**12.51**	**5.48**	**0.48**	**39.75**
Monocycloparaffins	0.00	0.00	0.00	0.00	0.00	1.53	7.89	7.38	3.37	0.30	20.48
Dicycloparaffins	0.00	0.00	0.00	0.00	0.01	4.45	5.81	4.08	1.60	0.16	16.11
Polycycloparaffins	0.00	0.00	0.00	0.00	0.11	0.49	0.99	1.05	0.51	0.02	3.16
Aromatics	**0.00**	**0.00**	**0.00**	**0.00**	**5.91**	**7.71**	**6.55**	**2.97**	**0.75**	**0.04**	**23.93**
MonoAromatics	**0.00**	**0.00**	**0.00**	**0.00**	**5.64**	**6.87**	**5.82**	**2.89**	**0.75**	**0.04**	**22.02**
Alkylbenzenes	0.00	0.00	0.00	0.00	4.64	3.94	2.32	1.31	0.37	0.02	12.59
Benzocycloalkanes	0.00	0.00	0.00	0.00	1.01	2.91	3.43	1.48	0.36	0.03	9.20
Benzodicycloalkanes	0.00	0.00	0.00	0.00	0.00	0.02	0.08	0.10	0.03	0.00	0.23
Diaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.26**	**0.74**	**0.72**	**0.09**	**0.00**	**0.00**	**1.82**
Naphthalenes	0.00	0.00	0.00	0.00	0.26	0.74	0.55	0.02	0.00	0.00	1.57
Biphenyls	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.07	0.00	0.00	0.25
Naphthocycloalkanes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Fluorenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Triaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Phenanthrenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenanthrocyclolalkn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic Sulfur	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.09**	**0.00**	**0.00**	**0.00**	**0.00**	**0.10**
Benzothiophenes	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.10
Dibenzothiophenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

4.3 Integration of PIONA and GC-FIMS results

Both PIONA and GC-FIMS were used to determine the hydrocarbon type distribution of the selected samples. It is necessary to integrate the results from both instruments to obtain a complete compositional profile. A computer program was created and the flow chart is shown in Figure 4.1. The program inputs the routine PIONA and GC-FIMS reports, then integrates the paraffins, naphthenes, and aromatics respectively. The integration of n-paraffins and iso-paraffins is straightforward since both PIONA and GC-FIMS separate n-paraffins and iso-paraffins in their #C distribution reports. Paraffins of #C≤11 come from PIONA, and the rest are from GC-FIMS.

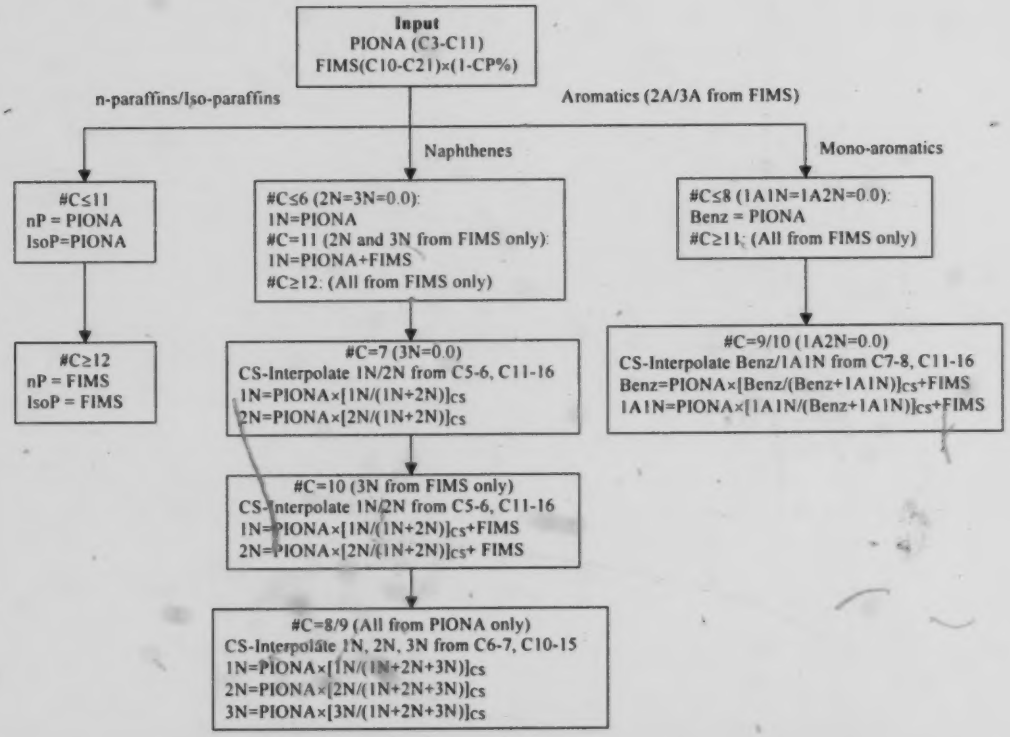


Figure 4.1. A Program Flow Chart for Module PIONA-FIMS Integration
Notations: CP%-cut point wt% at 200°C, CS-cubic spline, 1N-mononaphthenes, 2N-dinaphthenes, 3N-polynaphthenes, Benz-benzenes, 1A1N-benzomononaphthenes, 1A2N-benzodinaphthenes, 2A-diaromatics, 3A-triaromatics

However, the integration of naphthenes is complicated due to the lumping of naphthenes in PIONA and overlaps between PIONA and GC-FIMS results. There are three subgroups (monocycloparaffins, dicycloparaffin, and polycycloparaffins) in GC-FIMS data, while all naphthenes are lumped in PIONA results. At C10-C11, both PIONA and GC-FIMS could contribute to the naphthenic distribution. For the naphthenes with #C≤6 and #C≥11, explicit delumping of PIONA results and overlaps between PIONA and GC-FIMS can be determined from the chemical features and boiling behavior of each subgroup. The distributions of naphthenes between C7 and C10 are ambiguous due to the overlaps among the subgroups. The Gamma distribution function and Cubic Spline (CS)

interpolation have been attempted to estimate the 1N-3N subgroups from PIONA at the missing points (C7-C10), the distribution difference between two methods is negligible. For simplicity and flexibility, CS interpolation is employed in this work.

Generally, the following strategies are used for integrating naphthenes: a) For #C≤6, the results come from PIONA and only contribute to monocycloparaffins (1N). b) For #C≥12, the naphthenic distributions attribute to the GC-FIMS results, where all naphthenes boil above 200°C. c) At #C=11, 2Ns and 3Ns, normally boiling above 200°C, are from GC-FIMS measurements. Therefore, the distribution of 1N is obtained from the combination of PIONA and GC-FIMS results. d) At #C=7, 3N structures do not exist. CS interpolation is used to estimate the 1N and 2N distribution from the closest 8 known-points. The resulting distribution is normalized with the PIONA measurements. e) At #C=10, the distribution of 3Ns attributes to the GC-FIMS data only (boiling above 200°C), where the distributions of 1Ns and 2Ns are the sums of contributions from both measurements. Again, CS interpolations are used to split the contributions from PIONA to 1N and 2N subgroups. f) At #C=8/9, all naphthenes are from PIONA measurements. CS interpolations are applied to estimate the 1N-3N distribution from the closest 8 known-points. The resulting distributions are normalized to the PIONA yields.

The integration for aromatics is relatively simple since the PIONA results contribute to mono-aromatics only. Therefore, the following strategies are made for integrating aromatics: a) All aromatics rather than mono-aromatics are directly normalized from GC-FIMS data by (1-CP%). b) For #C≤8, PIONA results only go to alkyl-benzene distribution (benzo-cycloparaffins are unlikely to exist). c) For #C≥11, all mono-aromatics come from GC-FIMS results. d) At #C=9/10, benzo-dicycloparaffins (1A2N)

are not likely to exist. The distributions of alkyl-benzenes and benzo-cycloparaffins (1A1N) are the combinations of PIONA and GC-FIMS results. Similarly, CS interpolations are employed to split the contributions from PIONA to benzenes and 1A1N. The program inputs the PIONA (Table 4.2) and GC-FIMS (Table 4.3) Excel spreadsheets and then generates the integrated results for sample 1 in a separated spreadsheet as shown in Table 4.4. Similar integrations have been done for sample 2 through 5. The integrated results are tabulated in Appendix C, Tables C9-C12.

Table 4.4 Integrated results from PIONA and GC-FIMS measurements of sample 1

HC Type / #C	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	Sum
Saturates	**0.159**	**1.171**	**3.138**	**8.743**	**19.18**	**16.30**	**12.78**	**10.03**	**4.873**	**0.599**	**76.976**
affins	**0.057**	**0.455**	**1.524**	**4.798**	**13.93**	**11.02**	**6.802**	**4.934**	**2.643**	**0.404**	**46.561**
isoparaffins	0.023	0.171	0.739	2.047	8.391	6.197	4.450	3.429	2.270	0.386	28.102
n-Paraffins	0.034	0.284	0.785	2.752	5.537	4.821	2.351	1.505	0.373	0.018	18.459
Cycloparaffins	**0.102**	**0.716**	**1.615**	**3.945**	**5.256**	**5.284**	**5.977**	**5.093**	**2.230**	**0.196**	**30.415**
Monocycloparaffins	0.102	0.498	1.123	2.751	3.525	3.273	3.211	3.005	1.372	0.124	18.984
Dicycloparaffins	0.000	0.218	0.490	1.182	1.686	1.813	2.363	1.660	0.653	0.064	10.130
Polycycloparaffins	0.000	0.000	0.001	0.013	0.045	0.198	0.403	0.428	0.206	0.007	1.301
Aromatics	**0.011**	**0.250**	**1.740**	**8.095**	**5.589**	**3.138**	**2.665**	**1.211**	**0.306**	**0.017**	**23.022**
MonoAromatics	**0.011**	**0.250**	**1.740**	**8.095**	**5.480**	**2.797**	**2.370**	**1.176**	**0.305**	**0.017**	**22.242**
Alkylbenzenes	0.011	0.250	1.740	6.892	4.167	1.603	0.944	0.532	0.150	0.007	16.296
Benzocycloalkanes	0.000	0.000	0.000	1.204	1.313	1.185	1.395	0.601	0.145	0.010	5.853
Benzodicycloalkanes	0.000	0.000	0.000	0.000	0.000	0.009	0.031	0.043	0.011	0.000	0.094
Diaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.107**	**0.302**	**0.294**	**0.035**	**0.001**	**0.000**	**0.740**
Naphthalenes	0.000	0.000	0.000	0.000	0.107	0.302	0.222	0.007	0.000	0.000	0.638
Biphenyls	0.000	0.000	0.000	0.000	0.000	0.000	0.071	0.028	0.001	0.000	0.101
Naphthocycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001
Fluorenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Triaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Phenanthrenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phenanthrocyclolalkn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Aromatic Sulfur	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.038**	**0.000**	**0.000**	**0.000**	**0.000**	**0.040**
Benzothiophenes	0.000	0.000	0.000	0.000	0.001	0.038	0.000	0.000	0.000	0.000	0.040
Dibenzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

4.4 Data reconciliation between GC-FIMS and SimDis

As one of the validation checks, the final molecular representation derived from the GC-FIMS data was compared with the SimDis results. Prior to this comparison, the consistency between the GC-FIMS measurements and SimDis results had to be checked. This consistency check can be considered as data Quality Control (QC). Such a QC is very important to develop accurate molecular representations. The first QC step is to bring the two measurements to the same BP scale. In GC-FIMS, a HP5890 GC is coupled with a field ionization source and a MS detector, while SimDis uses a HP6890 GC and a FID detector. The columns, the injectors, and the heating procedures are quite different between the two methods. Therefore, each of them has its own paraffinic calibration standards (#C-RT-BP) to scale the retention time (RT) with BP. These standards are tested before each sequence of samples to ensure the measurements are run on the same BP scale. The reconciliation algorithm is shown in Figure 4.2. The program inputs #C-RT-BP standard tables, the chromatograms of paraffinic standards, and the chromatograms of real samples from SimDis and GC-FIMS respectively. Wätzig's method was applied to identify and sort the peaks (Wätzig, 1992). Using the n-paraffin #C-RT-BP standard tables and the CS interpolation method, the corresponding BPs referring to individual peaks are calibrated for each paraffinic standard tested on SimDis and GC-FIMS. At each #C, the calibrated BP is compared with the standard BP. If the difference is below a tolerance (e.g. tolerance =1.0°C in this work), the first QC check is passed and the real sample is tested after. Otherwise, the instruments need to be recalibrated.

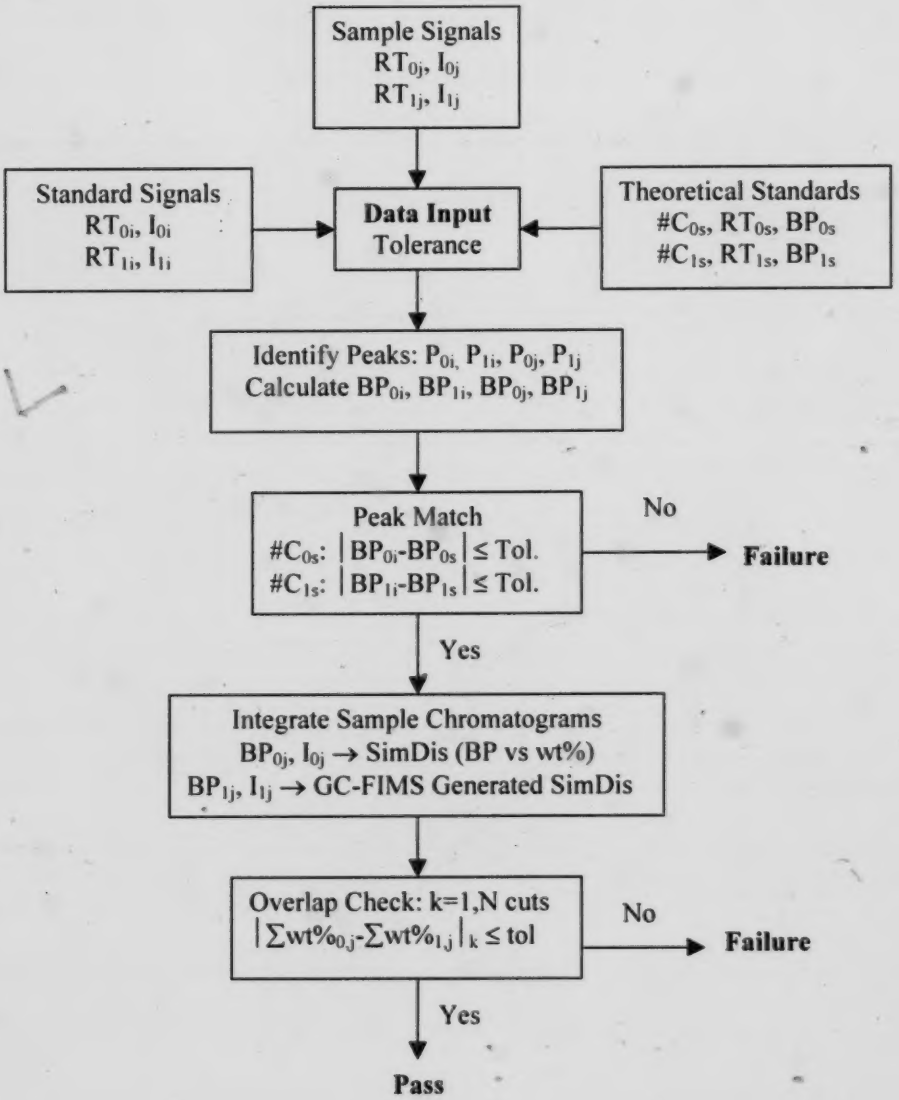


Figure 4.2 Flow Chart for Data Reconciliation between GC-FIMS (1) and SimDis (0)
Notations: I-Intensity, RT-Retention Time, BP-Boiling Point, P-Peak, Tol-Tolerance

The reconciliation results of this paraffinic calibration check for sample 1 are listed in Table 4.5. The calibrated BPs are consistent with n-paraffin BP standards with a maximum difference of 0.8°C (between SimDis and GC-FIMS). Such a QC check has also been applied to the rest of the samples (sample 2 to 5) before loading.

Table 4.5 Paraffinic calibration (RT-BP) check for SimDis and GC-FIMS

#C	BP-std	BP-SimDis	BP-GC/FIMS
5	36	35.963	36.764
6	69	69.054	68.560
7	98	97.984	98.109
8	126	126.021	126.100
9	151	151.041	150.992
10	174	174.039	173.825
11	196	196.037	196.155
12	216	216.078	215.611
14	253	253.037	253.459
15	271	271.015	271.134
16	287	287.014	287.101
17	302	302.035	302.047
18	317	317.036	317.273
20	344	344.036	344.653

In SimDis, a Flame Ionization Detector (FID) was used, in which the response factors were the same for all hydrocarbons. However, in GC-FIMS analysis, an MS detector was applied and different response factors were assigned to different hydrocarbon types. These sensitivity factors vary significantly with the compound classes (Briker et al., 2001a). The difference could be as large as 50 fold. Therefore, the chromatograms collected from these two measurements were not directly comparable unless these response factors were applied for GC-FIMS data. As illustrated in Figure 4.2, integrating the SimDis chromatogram results in the SimDis curve. Integrating the GC-FIMS chromatogram with the pre-calibrated sensitivities for each hydrocarbon type generates a similar SimDis curve — namely GC-FIMS-generated SimDis (FIMS-gen- SimDis). An overlap-check was conducted by comparing these two SimDis yields in every 10°C cut. If the average difference on each cut is below a certain tolerance (e.g. 1.0 wt%), the GC-FIMS test is considered to be consistent the SimDis result. Otherwise, repeating the experiment is necessary. The overlap-check result for sample 1 is shown in Figure 4.3. A good agreement has been found between the FIMS-gen-SimDis and the SimDis for

sample 1. The data reconciliation procedure is applied to the rest of the samples as well, and results are illustrated in Figure 4.4-4.7. The FIMS-generated simulated distillations agree with the measured SimDis curves very well, indicating a high accuracy in GC-FIMS measurements for all the selected samples. These QC checks ensure the consistency between the GC-FIMS report and SimDis results and generate reliable compositional analysis.

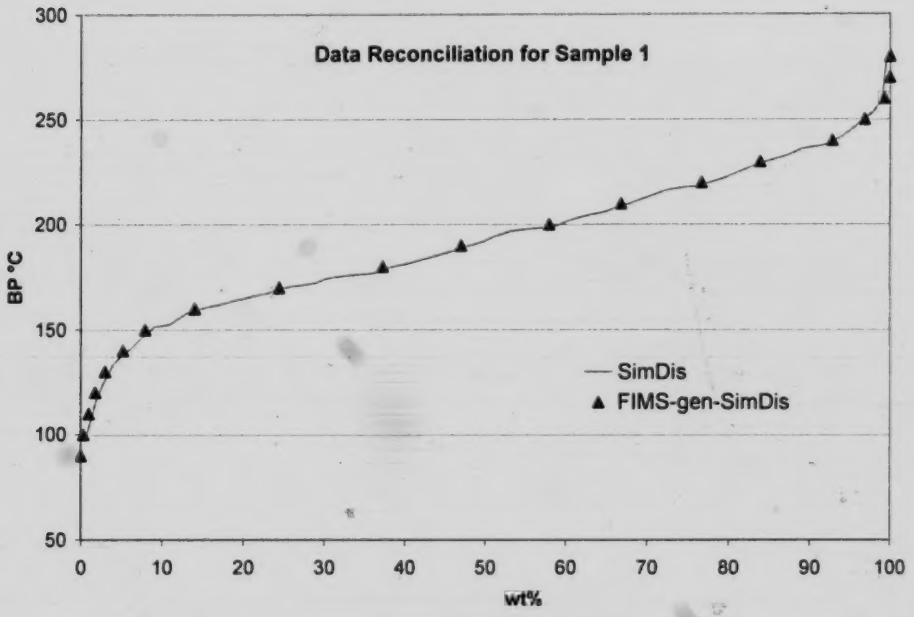


Figure 4-3. Data reconciliation result for sample 1

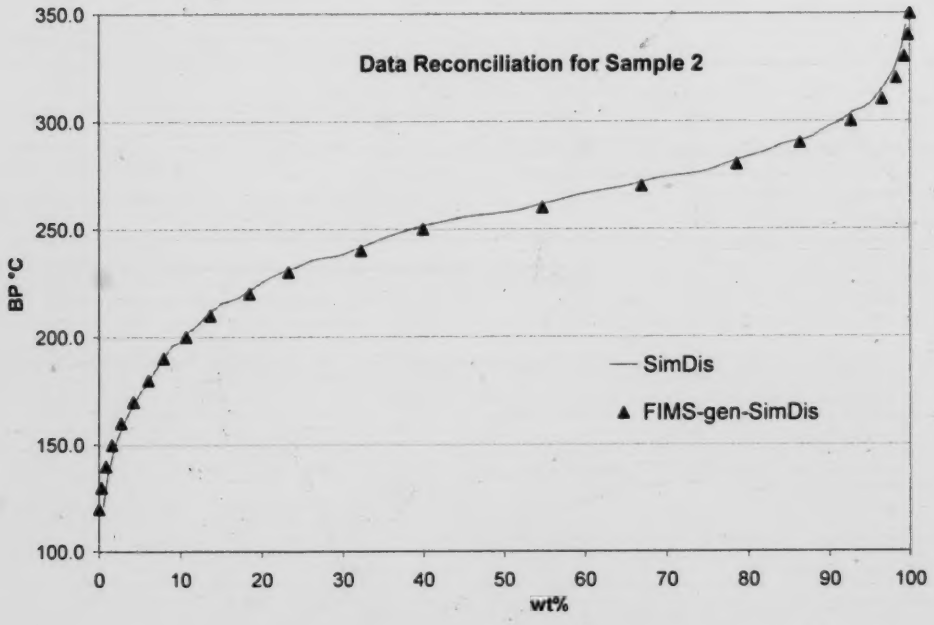


Figure 4-4. Data reconciliation result for sample 2

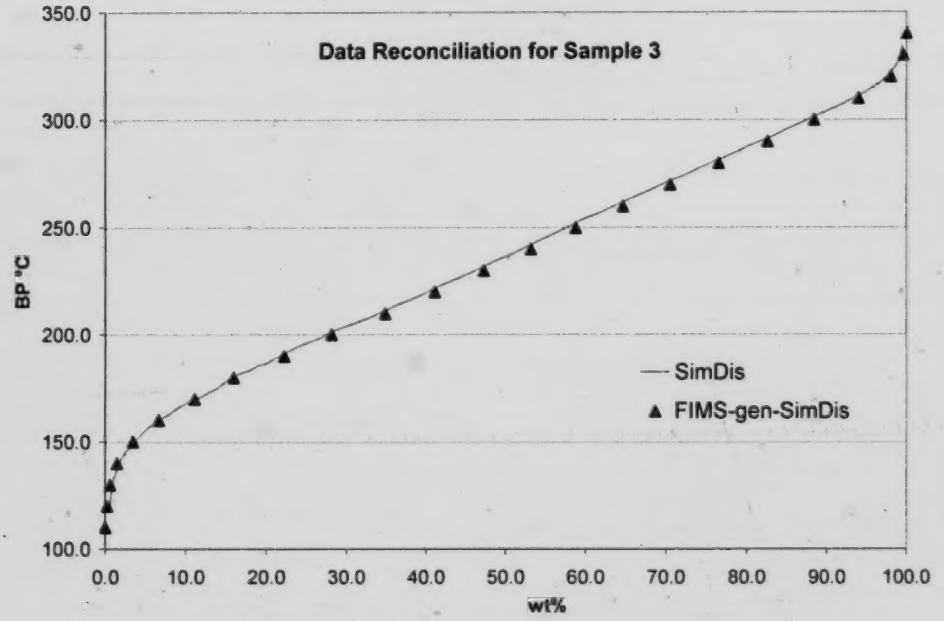


Figure 4-5. Data reconciliation result for sample 3

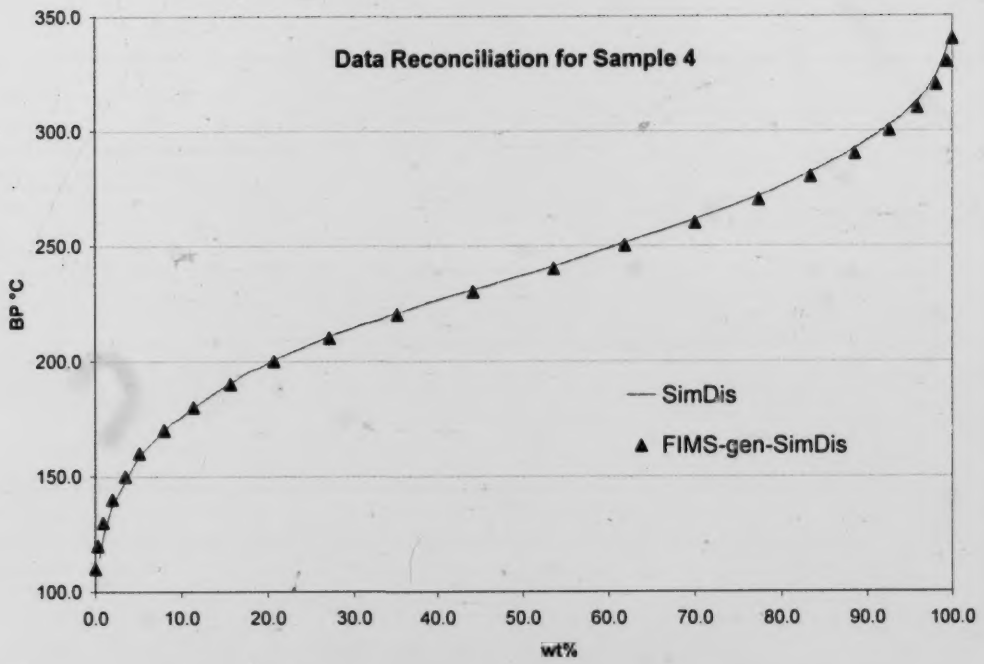


Figure 4-6. Data reconciliation result for sample 4

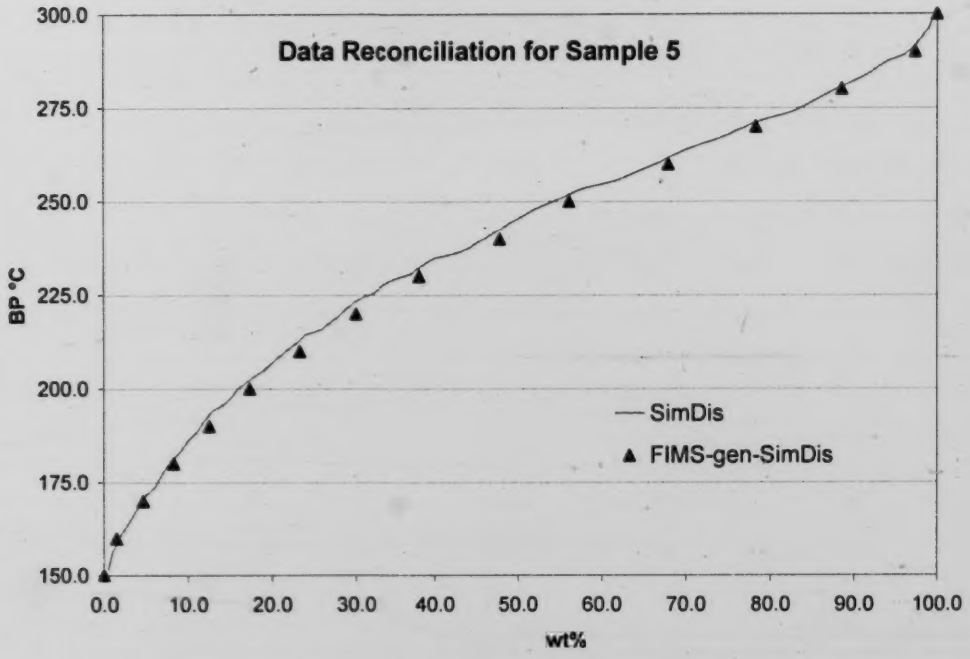


Figure 4-7. Data reconciliation result for sample 5

4.5 Conclusions

As a result of integrating PIONA and GC-FIMS results, accurate hydrocarbon type distribution by #C can be determined. The procedure avoids the uncertainties introduced by the possible fragmentation of the light ends of isoparaffins in GC-FIMS. As a means of QC checks, paraffinic calibration (#C-RT-BP) and overlap-checks have been applied to the selected samples (sample 1 to 5). The first step brings the two measurements (SimDis and GC-FIMS) to the same BP scale, and the latter guarantees the consistency between these two tests. The consistencies between the SimDis and FIMS-gen-SimDis through samples 1 to 5 indicate that the proposed data reconciliation procedures in this work are valid. The reconciliation ensures accurate analytical data from GC-FIMS, which provides a sound database to build the molecular representation. The next chapter will use the reconciled analytical data for detailed molecular characterization.

4.6 Reference

Altgelt, K. H.; and Boduszynski, M. M. *Composition and Analysis of Heavy Petroleum Fractions*, Marcel Dekker, Inc., 1994.

Annual Book of ASTM Standards, Vol. 5, ASTM, PA, 2001.

Boduszynski, M. M. (1988). Composition of heavy petroleums. 2. molecular characterization. *Energy & Fuels*, **2**(5), 597-613.

Briker, Y.; Ring, Z.; Iacchelli, A.; McLean, N.; Rahimi, P. M.; and Fairbridge, C. (2001[a]). Diesel fuel analysis by GC-FIMS: aromatics, n-paraffins, and isoparaffins. *Energy and Fuel*, **15**, 23-37.

Briker, Y.; Ring, Z.; Iacchelli, A.; McLean, N.; and Fairbridge, C. (2001[b]). Diesel fuel analysis by GC-FIMS: normal paraffins, isoparaffins, and cycloparaffins. *Energy and Fuel*, **15**, 996-1002.

Chasey, K. L.; Aczel, T. (1991). Polycyclic aromatic structure distributions by high-resolution mass spectrometry, *Energy & Fuels*, **5**, 386-394.

Neurock, M. N.; Nigam, A.; Trauth, D.; Klein, M. T. (1994). Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms, *Chem. Eng. Sci.* **49**, 4153-4177.

Sheremata, J. M.; Gray, M. R.; Dettman, H. D.; and McCaffrey, W. C. (2004). Quantitative Molecular Representation and Sequential Optimization of Athabasca Asphaltenes, *Energy & Fuels*, **18**, 1377-1384.

Wätzig, H. (1992). Human Judgment, *Chromatographia*, **33**, 218-224.

Chapter 5

Derivation of Molecular Representations of Diesels

5.1 Introduction

Accurate characterization of these feedstocks plays an important role in kinetic modeling and integrated process optimizations. In refining conversion processes such as hydrocracking and fluid catalytic cracking (FCC), the reactivity and product yield distributions strongly depend on the molecular composition of feedstocks. Feedstocks of different molecular compositions have different cracking patterns. Advances in the characterization of petroleum fractions are closely related to the development of new more sensitive and less expensive analytical techniques. The composition of light petroleum fractions (i.e. naphtha) can be measured at the molecular level using the DHA (Detailed Hydrocarbon Analysis) method. However, compositions of middle distillates and vacuum gas oils that are particularly difficult to characterize can only be measured in terms of hydrocarbon class distributions by boiling point (Boduszynski, 1988), or by carbon number (Bouquet and Brument, 1990; Chasey and Aczel, 1991). One of the analytical techniques particularly useful for middle and heavy distillates is gas-chromatography field ionization mass spectrometry (GC-FIMS).

Consistent with the state-of-the-art in analytical technology, petroleum mixtures are characterized in terms of several molecular classes (homologous series) distributed by carbon number in the development of mechanistic kinetic models, such as the Structural Oriented Lumping (SOL) model (Quann and Jaffe, 1992) and the Single-Event Kinetic model (Hillewaert et al., 1988). A strong assumption for lumping a petroleum mixture into homologous series is that the physical and chemical properties of the hydrocarbon

isomers are identical to those of the defined species from molecular classes. Clearly, this assumption is not true for most of the hydrocarbons, considering the intractable numbers of isomers beyond C_{10}. Many thermo-physical properties of the isomeric lump are not reliable because of the wide spread of values of corresponding properties of isomers in the lump. Some thermo-physical properties of isomers can vary widely according to their molecular structures (API technical data book, 1992). For hydrocarbon compounds, the catalytic reaction mechanisms can be quite different among isomers. The cracking of three C_6 isomers (2-methylpentane, 3-methylpentane, 2,3-dimethylbutane) on USHY shows that these three undertake quite different reaction paths and, thus, yield different cracked product distributions (Wojciechowski, 1998). The molecular structure information is crucial to the reactivity of hydrocarbons. The reaction rates for various isomers can be quite different, especially where shape-selective catalysts are involved (Krambeck, 1991). In mechanistic kinetic models, the reactivity is often correlated with reaction indices that depend on the structure of compounds. Therefore, having the detailed distribution of the isomers within an isomeric lump is important for process modeling, especially for kinetic modeling.

Neurock et al. (1994) developed a Monte Carlo simulation method to characterize the complex mixtures. In this method, bulk properties from elemental analysis and NMR measurements were used to build the probability curves for various structural attributes of hydrocarbon molecules, such as side chains, aromatic rings, and naphthenic rings. The Monte Carlo technique was then used to randomly sample the probability distributions and construct the molecules, thereby forming a set of molecules to fit the analytical bulk data. Molecules in the order of 10^5 were sampled to represent the petroleum mixtures

including heavy oil, vacuum gas oil, and asphaltenes. However, the completely random sampling easily introduces redundant molecules. The randomly generated molecule does not reflect the thermodynamic features of the molecule that play a key role in its stability. In addition, the current computational technology cannot afford this large ensemble of molecules in kinetic modeling. Efforts had been made to derive a small set (10-100) of molecules from Monte Carlo representations using quadrature method (Campbell and Klein, 1997). The mole fractions of this small set are further optimized to match experimental bulk data.

However, molecular composition is only partially reflected by bulk properties such as gravity, elemental content, and NMR analysis. Kim et al. (1998) investigated the compositions of three FCC feedstocks that had very similar bulk properties. Through mass spectrometry they found that the hydrocarbon class distributions in those feedstocks were significantly different. Ramaswamy et al. (1989) also found that the hydrocarbon class distributions of two vacuum gas oils were entirely different although the two samples had almost the same results from SARA (Saturates, Aromatics, Resins and Asphaltenes) analysis. Therefore, molecules completely constructed from the average structural parameters may lead to an uncertainty in the composition of petroleum mixtures; as a result, the product yields and properties might be totally different even though such feedstocks have very similar bulk properties. For those heavy barrel bottoms, like residue and asphaltenes, only the bulk properties can be measured and the stochastic characterization methods might be the only choice (e.g., Neurock et al., 1994; Sheremata, et al., 2004). However, for the petroleum fractions like diesels or gas oils, the detailed hydrocarbon class distribution can be determined by GC/LC-MS. Accurate molecular

representations should be based on these detailed compositional information if they are available. The objective of this work was to develop a deterministic method for molecular characterization of petroleum middle distillates. This chapter uses the GC-FIMS determined hydrocarbon class distribution as the starting point, generates molecules from a set of construction rules, de-lumps each measured isomeric lump into a set of isomers most probable to exist, and presents a detailed molecular representation for diesels. Developing such a deterministic method enables a fast and reliable molecular representation for petroleum middle distillates.

5.2 Methodology

Starting from a reconciled "characterization matrix" measured by GC-FIMS, which determines 19 hydrocarbon homologous series distributed by carbon number (#C), a delumping algorithm is proposed to derive a molecular representation of diesels (see Figure 5.1). For each measured isomeric lump, a group of isomers is generated from a set of molecular generation rules. The generated molecules are optimized geometrically on-line using MOPAC PM3 procedure and a subset of isomers is selected using a suitable heat of formation threshold. For each selected molecule, boiling point (BP), specific gravity (SG), and refractive index (RI) are predicted from QSPR correlations developed by Ha et al. (2005a). Thermodynamic properties like free energy of formation are also calculated for each molecule using MOPAC. With the set of representative isomers at hand, the distribution of isomers with an isomeric lump can be determined by minimizing the Gibbs free energy of the lump subject to the stoichiometric constraint and the measured average boiling point of that lump (Ha et al., 2005b). Using a proper value of

molecular pre-screening threshold (2.0 kcal/mol in heat of formation used in this work), an adequate molecular set is obtained, from which good agreements can be found between measured bulk properties and the predicted ones using mixing rules, and thus, validates the methodology proposed. A minimum set of representative molecules is then sorted out by neglecting the concentrationally insignificant molecules without affecting the agreements between the measured and predicted bulk properties (SimDis, density, and RI). In general, the methodology adopted in this work consists of following steps:

1) Analytical measurements and data reconciliation
2) Molecular generation of isomers and on-line simulation
3) QSPR predictions for physical properties and thermodynamic calculations
4) Isomer distribution within each measured isomeric lump
5) Molecular refining to derive the minimum set of molecules

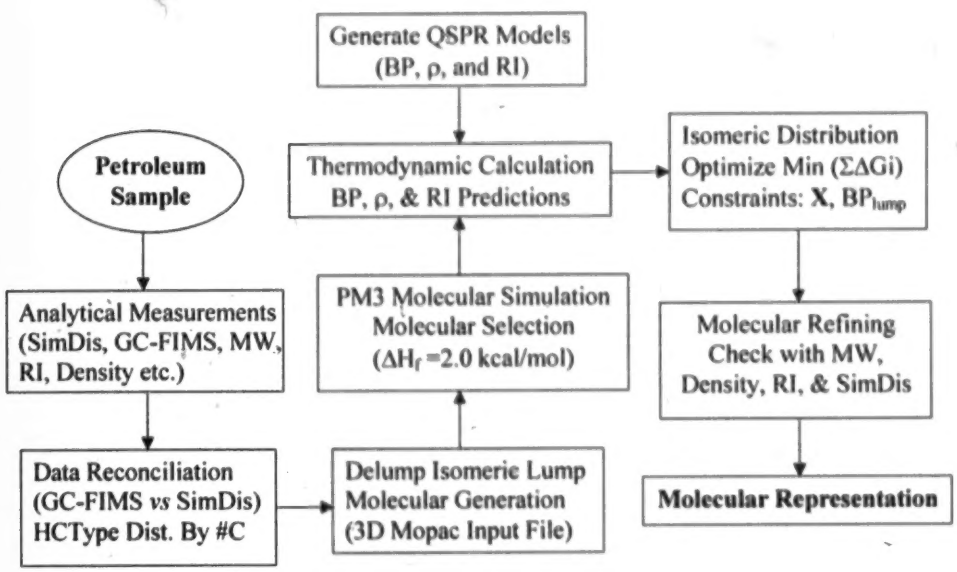


Figure 5-1. Molecular delumping algorithm for molecular characterization

5.2.1 Analytical measurements and data reconciliation

As introduced in Chapter 4, five compositionally different diesel samples have been used for testing the characterization methodology. The compositional details and bulk properties of the samples are summarized in Table 4-1. PIONA and GC-FIMS tests were done for the selected samples. An integration procedure was applied to integrating both the PIONA and GC-FIMS results and obtaining a complete range of hydrocarbon type distributions for diesels. Those results are tabulated in Table 4-2 to 4-4, as well as in Appendix C, Tables C1-C12. To ensure accuracy, the integrated GC-FIMS results have been converted into an equivalent SimDis curve and compared with the experimental SimDis results. This data reconciliation procedure was applied to all the samples. The molecular characterization does not proceed until a consistency was found between the GC-FIMS report and SimDis results. As shown in Figures 4-3 to 4-7, the GC-FIMS generated SimDis agrees with the measured SimDis curve very well for all selected samples, indicating a high accuracy in GC-FIMS measurements.

5.2.2. Molecular generation and simulation

5.2.2.1. Molecular generation rules

In stochastic methods where analytical data were limited to bulk properties, the molecules used for this representation were randomly generated from structural building blocks (Neurock et al., 1994). Then a large assemble of molecules were optimized with the derived distribution functions of structural attributes until a set of bulk physical properties of the mixture is sufficiently close to the measured values. In this work, the detailed GC-FIMS compositional matrix (in isomeric lumps) is available and a

deterministic method is explored for generating molecules to avoid the redundancy and improve the generation efficiency. The way to build molecules is based on the reported molecular occurrences in petroleum, and simulation results of those molecules that are difficult to isolate and quantify. Each cell of the by-#C characterization matrix was assumed to be an isomeric lump for a specific hydrocarbon class at a given #C. The number of possible isomers within the lump increases rapidly with the number of carbon atoms. Even for light distillates (#C<21), the number of all possible isomers for a specific #C can be in millions (Read, 1976). However, not all isomers are equally probable to be present in the isomeric lump in significant concentrations. The occurrences of individual isomers depend on their origin, maturing process, and thermodynamic stability (Tissot and Welte, 1984). The molecular generation rules are derived from following observations in open literature and molecular simulations of this work.

For alkanes, the 2- and 3-methyl derivatives are the most abundant, and the 4-methyl derivative is present in small amounts. It is generally accepted that the slightly branched paraffins predominate over the highly branched materials (Speight, 1999). Tissot and Welte (1984) concluded that the most frequent configuration of paraffins is one tertiary carbon atom (2-methyl or 3-methyl). The configuration with two tertiary carbon atoms is less abundant. Other types (one quaternary carbon atom, or more than two tertiary atoms) are usually very rare. The small alkanes are not inherited biogenic molecules, but are generated through thermal degradation and cracking of C-C bonds of either kerogen or larger bitumen molecules already formed. The compositional investigations of 18 crude oils by Martin et al. (1963) also showed the prevailing occurrences of n-paraffins and 2/3-methyl branched paraffins in naphthas of those crudes. Obviously the nature favors

the simple structures of paraffins, at least within the diesel range. However, the thermodynamic data (both heat of formation and free energy of formation) of paraffin isomers in naphtha show that the 2,2-dimethyl branched paraffins are the most thermodynamically stable ones, while n-paraffins and 2/3-methyl branched paraffins were less stable than the ones with two methyls on main chain far-apart. For example, Figure 5.2 illustrates the reported free energy of formations for C9 isoparaffins (API Technical Data Book, 1992). In general, the paraffin molecules with just one or two methyl branches are more stable than those having more than three branches or having branches longer than ethyl.

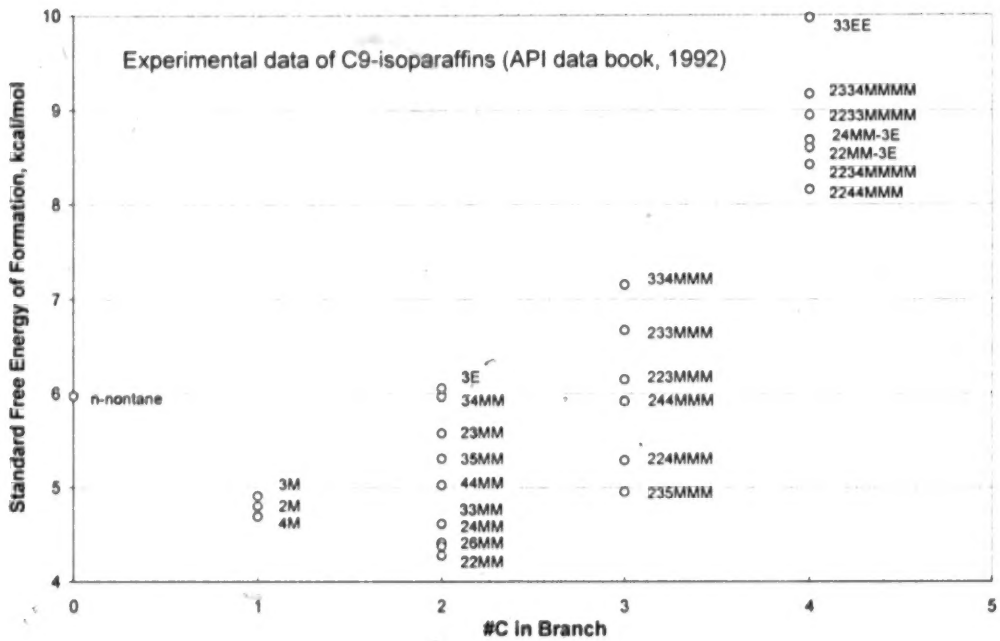


Figure 5.2 Thermodynamic stabilities of C9 isoparaffins (M-methyl, E-ethyl)

The NMR analyses for crude oils and their distillates showed the tertiary aliphatic carbon accounts for about 10% of total carbon and the branch index (CH_3/CH_2) of

saturates range from 0.1 to 0.3 (Japanwala et al., 2002; Ramaswamy et al, 1989; and Khan et al, 2000). Considering the maximum #C in GC-FIMS characterization matrix is 21, we limit the isoparaffins in diesel to have at most three branches with a maximum length of C4 and no quaternary carbon is allowed. Similar rules were made by Mizan and Klein (1999) in their study of catalytic hydroisomerization to rationalize the product spectrum.

Based on the fact that only cycloparaffins with five- and six-membered rings have been isolated from the petroleum light fractions, and that thermodynamic studies show naphthenic rings with five and six carbon atoms are the most stable (Speight, 1999), it can be assumed that only these cycloparaffins are present. Methyl-cyclohexanes and methyl-cyclopentanes are frequently the most abundant in the light fractions (#C<10). Mono- and Dicycloalkanes generally account for half of the total cycloparaffins for #C>10. Tricycloalkanes are most likely to have a structural arrangement similar to perhydrophenanthrene. The abundance of tricycloalkanes decreases with #C beyond 20. The structures of tetra- and pentacycloalkanes are directly related to the tetracyclic steroids and pentacyclic triterpenes (Tissot and Welte, 1984). In general, tetra- and pentacycloalkanes are most abundant in immature crude oils. They are beyond the range of diesels and reported in gas oils.

The aromatics in crude oils usually include one to five condensed aromatic rings with a small number of short chains. Among them, alkylated homologues of benzene, naphthalene, and phenanthrene are the most abundant (Tissot and Welte, 1984). They often include one to three carbon atoms in substituent chains. For instance, the most abundant alkyl-benzene in the gasoline fraction is frequently toluene or xylene, whereas

benzene is usually less abundant. The same is true for naphthalene and phenanthrene, where ethyl- or propyl- substituted naphthalenes and phenanthrenes are the most abundant. When several structural arrangements are possible, as for molecules with three or more aromatic rings, only a few of them are favored in crude oils. For example, alkyl-phenanthrenes are largely predominant triaromatics. Among the five isomeric configurations of tetra-aromatics, cata-condensed chrysene is likely to be the most prevalent form, based on the origin of ring structures in petroleum (Tissot and Welte, 1984). Naphthenoaromatics are particularly abundant, as compared to pure aromatics, in shallow immature crude oils. The aromatic types become dominant after a significant thermal evolution. Bicyclic indane, tetralin, and their methylderivatives are usually abundant. Tricyclic tetrahydro-phenanthrene and derivatives are also quite common. Tetracyclic and pentacyclic molecules, which are mostly related to steroid and triterpenoid structures with 1-3 aromatic rings, are important in the high boiling fractions like vacuum gas oils (Tissot and Welte, 1984).

The distribution of alkyl groups may vary with the number of groups on a ring, the carbon number of the substituents, and the branching structure of the groups. With the NMR measurements on a physically isolated monoaromatic fraction of petroleum, Mair and Barnewall (1964) deduced that, on average, alkyl benzenes in the C_{13}-C_{15} range had one methyl substituent and one longer chain with a methyl branch. By comparing the MS fragmentation pattern of both synthetic and petroleum alkylbenzenes in the C_{20}-C_{40} range, Hood et al. (1959) concluded that the alkyl group structure is not as diverse as suggested by the number of possible isomers. They observed that petroleum alkylbenzenes are

composed almost entirely of those having a single long chain and from 0 to 4 methyl substituents on the benzene ring. It is interesting that the distribution of substituents differs between aromatics and naphthenic rings within the same molecule. The short chains (methyl and ethyl) appear to be prevalent substituents of the aromatic portion of the ring cluster, whereas a limited number (one or two) of longer chains may be attached to the naphthenic rings. The total number of chains, which is in general four to six, as well as their lengths, increases with #C (Hood et al, 1959).

A molecular simulation study of the thermodynamic stabilities of various aromatic isomers also showed a tendency to a higher degree of substitution in alkylated aromatics. Figure 5.3 illustrates the standard heat of formation of C10 alkyl-benzenes, where simulated results are consistent with the experimental ones in terms of the relative thermodynamic stabilities. The heat of formation is used as an approximate index of thermodynamic stability since it accounts for the dominant part of free energy of formation of hydrocarbons under moderate conditions. Figure 5.2 and 5.3 also show that with the same number of branches, the molecules with their branches well separated each other are more thermodynamically stable than those having crowded branch distributions. For instance, 2,6-dimethyl-heptane (except for 2,2-dimethyl-heptane) and 1,3-dimethyl-5-ethyl-benzene are the most stable structures for di-methyl-heptanes and tri-branched C10-benzenes, as showed in Figure 5.2 and 5.3, respectively. Based on the above observations and molecular simulations, the following molecular generation rules were applied in this work:

a) There are maximum three branches in isoparaffins.

b) The maximum length of an isoparaffin branch is C4.

c) Quaternary C is not allowed in saturates.

d) There are maximum four substituents on a ring-core.

e) Only one non-methyl branch is allowed in the case of multiple substitution.

f) Multiple substituents are located on the main structure (ring or chain) as far apart as possible.

g) Only five- or six-membered rings are allowed in naphthenes.

h) The only non-methyl branch is put on the naphthenic rings of Naphthenoaromatics.

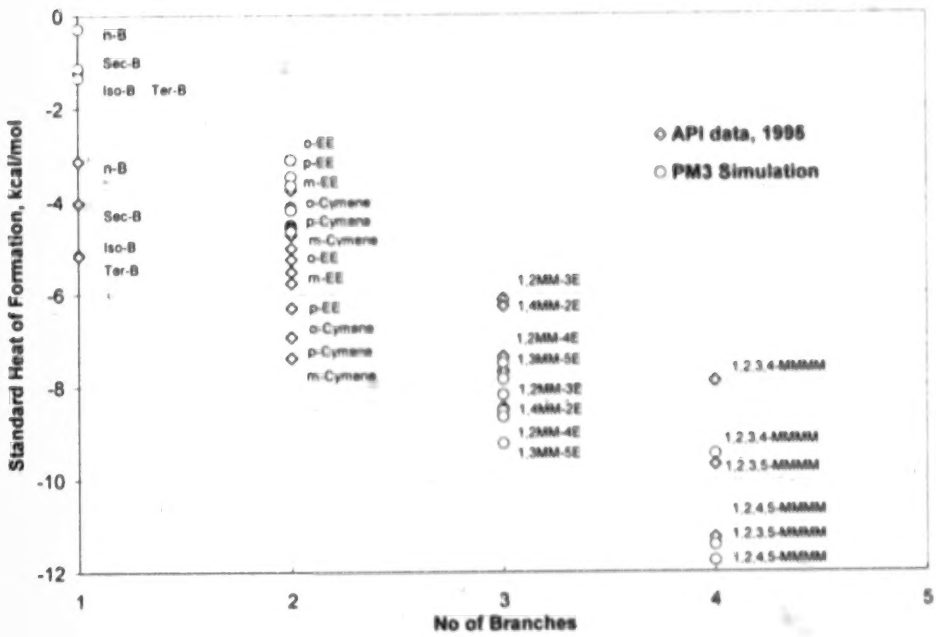


Figure 5-3. Molecular simulations of C10 alkyl-benzenes (B-butyl, M-methyl, E-ethyl)

5.2.2.2 Molecular fingerprint in MOPAC input format

Following the molecular generation rules, hydrocarbon molecules are constructed in the MOPAC input format. Figure 5.4 shows an example of a MOPAC input file for

ethane. The first line includes the functional key words, from which the calculation task is loaded. Lines 2 and 3 are comments. Line 4 and afterwards specify the atoms, the bond length, bond angle, and dihedral angle among the atoms, with an optimization flag (0 or 1) following behind. The MOPAC input file offers three new features in molecular characterization: presenting a detailed three dimension description of molecular structure, enabling the molecular view (using Chem3D etc.), and allowing on-line molecular simulation. The generated MOPAC input uses internal coordinate representation and provides an initial estimate of the 3D conformation of molecules. Broadbelt et al. (1994b) proposed an algorithm for translating a 2D molecular graph into a 3D internal coordinate representation of MOPAC input. A similar algorithm is used in this work; however, instead molecules are directly assembled from their main structures (chain or ring core) and substituents. The bond length is determined from the connecting atoms and types of bond between them. For example, the bond lengths between two carbons are 1.52 and 1.39 Å for a single bond and a double bond respectively, whereas the bond length between carbon and hydrogen is 1.113 Å. The bond angle depends on the hybridization of parent atom. The values of 180°, 120°, and 109° are assigned for *sp*, sp^2, and sp^3 hybridizations respectively. For instance, the hybridization of carbon atom in methane is sp^3, therefore the bond angle of hydrogen atoms connected to carbon is 109°. The specification of dihedral angle is related to the types of parent, angle, and dihedral atoms and the order of these atoms being visited. Detailed specification method was described by Broadbelt et al. (1994b). Generally, an appropriate dihedral angle ensures the current atom be separated from others beyond a certain distance to avoid the overlap between any atoms.

PM3 NOMM NOXYZ NODIIS GRAPH T=10D

Atom (I)	Bond Length NA:I		Bond Angle NB:NA:I		Dihedral Angle NC:NB:NA:I		Parent Bond NA	Dihedral NB	NC
C(1)	0.000000	0	0.000000	0	0.000000	0	0	0	0
C(2)	1.540000	1	0.000000	0	0.000000	0	1	0	0
H(3)	1.090000	1	109.500000	1	0.000000	0	1	2	0
H(4)	1.090000	1	109.500000	1	120.000000	1	1	2	3
H(5)	1.090000	1	109.500000	1	-120.000000	1	1	2	3
H(6)	1.090000	1	109.500000	1	180.000000	1	2	1	3
H(7)	1.090000	1	109.500000	1	60.000000	1	2	1	3
H(8)	1.090000	1	109.500000	1	-60.000000	1	2	1	3

Figure 5-4. The internal coordinates (MOPAC input format) of Ethane

5.2.2.3 Molecular simulation, selection, and thermodynamic calculation

Semi-empirical calculations, like AM1 and PM3 procedures, provide a balance of computational efficiency and reliable estimates of a wide range of electronic and thermodynamic properties. They have been successfully used in the description of organic chemistry. Using PM3 procedure, Stewart (1990) reported that the accuracy of heat of formations for all organic compounds has a signed error of +0.21 kcal/mol. Therefore, PM3 procedure is applied in the geometry optimization of molecules. By definition, ΔH_f^0 is the calculated gas-phase heat of formation at 298K of one mole of a compound from its elements in their standard state. The ΔH_f^0 is used for molecular prescreening and for estimation of free energy of formation.

The number of molecules generated from the rules listed above is still very large. For instance, hundreds of isomers can be easily built for C20 cyclic hydrocarbons. To limit the number of molecules to a manageable level, a selection algorithm is required to screen the unimportant or unstable ones out. Representative molecules will be selected based on the thermodynamic stabilities of isomers, namely comparing the standard heat of formations generated during the PM3 optimization. The molecule with the lowest heat

of formation is taken as a reference. Molecules, whose standard heat of formation is within a threshold as compared with the reference molecule, are considered to be stable and selected as candidates to share the concentration of the isomeric lump measured by GC-FIMS. The threshold is an adjustable parameter, which controls the total number of representative molecules. 2.0 kcal/mol is used as the initial value for the threshold. At standard state (1atm, 298K), heat of formation is a dominant part of free energy of formation for hydrocarbons (Wade Jr, 1987). A threshold of 2.0 kcal/mol is equivalent to the same amount of difference in free energy of formation. Based on the thermodynamic calculations, a 2.0 kcal/mol difference in free energy of formation between two isomers yields a prevalent distribution (94/6) of one over the other (Wade Jr, 1987). Therefore, using 2.0 kcal/mol as the initial threshold will screen out those molecules thermodynamically unimportant. The thermodynamic calculations in MOPAC compute the vibrational frequency of each atom in a molecule, thus demanding extensive CPU time. Employing the prescreening criterion (ΔH_f = 2.0 kcal/mol) in the molecular generations will also significantly reduce the computational loads for thermodynamic calculations and thus improve the computational efficiency during molecular simulations.

For each selected molecule, its optimized geometry is updated and thermodynamic properties are calculated using MOPAC2002. Various thermodynamic quantities (partition function, enthalpy, heat capacity, and entropy) can be calculated from the vibrational frequency of the molecule. These quantities can be decomposed into vibrational, rotational, internal, and translational contributions. From these contributions, MOPAC calculated heat of formation and entropy refer to ideal gases at a prescribed

temperature. Using these results, the standard Gibbs free energy of formation of a selected molecule in gas phase can be derived (Stull et al., 1969).

$$\Delta G_f^o = \Delta H_f^o - T(S^o - n_C S_C^o - \frac{n_H}{2} S_{H_2}^o) \qquad (5\text{-}1)$$

Where n_C and n_H are the atoms of carbon and hydrogen, S_C^o = 1.361 cal/K.mol, the entropy of graphite at 298.15K, $S_{H_2}^o$ =31.211 cal/K.mol, the entropy of hydrogen gas at 298.15K. The ΔG^0_f will be used when isomers are distributed within an isomeric lump.

5.2.2.4 Generations and manipulations of hydrocarbon molecules

A molecular generation program, named module "MolGen", was created to construct the molecules in MOPAC input format for all the hydrocarbon types identified in GC-FIMS. A flow chart for coding is illustrated in Figure 5-5. As mentioned in the chapter 4, there are 19 hydrocarbon types (HCType) identified in a GC-FIMS measurement, which includes 3 aromatic sulfur types (benzothiophenes, dibenzophiophenes, and benzonaphthothiophenes). Since this work focuses on the molecular representation of diesels, the generation of sulfur compounds is not included here (the mount of sulfur compounds in diesel is negligible). As shown in Figure 5-5, the hydrocarbon molecules are generated from paraffins (Z=2) to tetra-aromatics (Z=-24), one-by-one in hydrocarbon type (or by Z decreasing on even numbers), over the whole boiling range of diesels (#C=5 to 21). Here, 16 hydrocarbon types are sorted as: nP (normal paraffin), iP (iso-paraffin), 1N (mono-cycloalkane), 2N (dicycloalkane), 3N (tricycloalkane), 1A (alkylbenzene), 1A1N (benzo-cycloalkane), 1A2N (benzo-dicycloalkane), 2A

(naphthalene), A_A (biphenyl), 2A1N (naphtho-cycloalkane), ANA (Fluorene), 3A (phenanthrene), 3A1N (phenanthro-cycloalkane), 4Ap (Pyrene), and 4Ac (Chrysene).

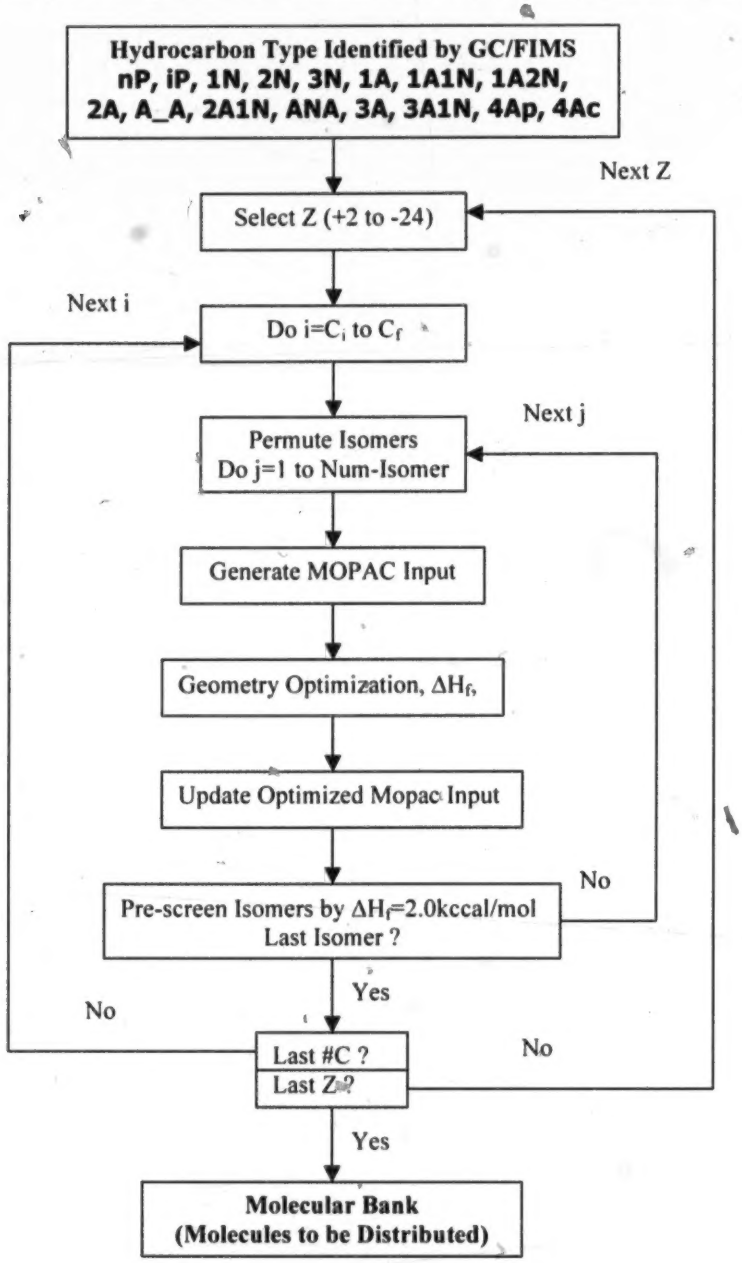


Figure 5-5. Programming diagram for molecular generation

Following this molecular generation scenario, a permutation subroutine is called to permute all the possible isomers for a selected hydrocarbon type and a specific #C, based on the main structures of molecules and the above molecular generation rules. The permuted isomers are stored in an isomer identification matrix (IsomerID) from which the MOPAC input of each individual molecule is generated by calling subroutines corresponding to each hydrocarbon type. The details of these subroutines will be introduced shortly. With the constructed molecule, "MOPAC2002" is called on-line to calculate the best 3D conformation of the molecule using PM3 procedures. The optimized 3D geometry is then updated to the initial MOPAC input used for later thermodynamic calculations. The PM3 optimization procedure calculates the standard heat of formation (ΔH^0_f) for the best conformation of the molecule. The ΔH^0_f is compared with the isomer having the smallest ΔH^0_f. If the difference in ΔH^0_f is larger than a threshold, the new generated molecule will be discarded, otherwise it will be selected as one of the representative isomers in that isomeric lump. This pre-screen procedure eliminates those thermodynamically unfavorable molecules, thus improve the computational efficiency for later thermodynamic calculations. The molecular generation proceeds until the last isomer in IsomerID is generated. Then the program moves to the next #C (isomeric lump) for another set of isomers until the final #C is reached. At this point, the molecular generation for a selected hydrocarbon type (Z series) is finished. The program then selects another hydrocarbon type, repeating the above molecular generation procedures until the last Z series is done. Finally, all the representative molecules are generated and selected with the optimized (updated) geometries. These molecules are stored in a molecular bank for isomer distribution and further manipulation.

5.2.2.4.1 Molecular generations of paraffins

Generating paraffins is schematically different from building the cyclic hydrocarbon molecules. Writing a MOPAC input for an n-paraffin molecule (HCType =1) is straightforward. Once the molecular size (#C) is known, the single n-paraffin molecule can be constructed linearly by adding carbon atoms one after the other with the corresponding hydrogen atoms attached on, like the MOPAC input file for ethane as shown in Figure 5.4. Schematic diagrams for building the isoparaffins are shown in Figure 5.6. At a given #C, isoparaffin isomers (HCType =2) are assembled by adding the branches on a straight main chain. The MOPAC input of the main chain is exactly the same as n-parafins, except with the Main Chain Carbon Number (MCCN). Based on the molecular generation rules, the longest branch for isoparaffins is C4, whereas longer substitutes are allowed for cyclic hydrocarbons. Different types of branches are also allowed for the substitutes having length of branch (LBr) longer than C3. The n-propyl and iso-propyl are allowed for LBr = 3, while 4 different butyls (n-, iso-, sec- and ter-) are used for LBr =4. The branch types used in the molecular construction are listed in Table 5.1. MOPAC input characteristics are generated separately for these different types of branches, which will be used for the overall molecular construction.

Table 5.1 Branch types being attached on main chain or main ring core

#C of Branch	Type 1	Type 2	Type 3	Type 4
1	methyl	/	/	/
2	ethyl			
3	n-propyl	iso-propyl		
4	n-butyl	iso-butyl	sec-butyl	ter-butyl
m	n-C_m-*	iso-C_m-*		

*n-C_m- and iso-C_m- represent a linear and an iso-branched-linear substitutes having *m* carbon atoms

As shown in Figure 5.6, iso-paraffin isomers are permuted first for a given #C. For those iso-paraffins having limited isomers (#C ≤ 8), the complete set of isomers are picked directly and assigned into an isomer identification matrix (IsomerID); e.g., 2, 4, and 8 isomers (excluding n-paraffin) are put into the IsomerID directly for C5-, C6-, and C7-isoparaffins respectively. The IsomerID is a 2D matrix with dimension of n×12, where n accounts for the number of isomer permuted. The 12 elements in the vector IsomerID(i,j=1 to 12) for isomer i represent the #C, MCCN, number of branches (#Br), lengths of 3 branches (3LBrs), types of 3 branches (3TBrs), and positions of 3 branches (3PBrs), accordingly. For the isoparaffins having #C ≥ 9, the isomer will be permuted by programming. With the constraints from the molecular generation rules (max #Br =3, max LBr =4, no quaternary carbon, only one non-methyl branch is allowed), isoparaffins having different #Br, LBr, TBr, and PBr are permuted and the results are assigned in the IsomerID matrix. The permutations proceed until all isomers under constraints are sorted out.

From the information of IsomerID, the main chain and branches are generated separately in the format of MOPAC input. Then the branches are inserted into the main chain MOPAC input one by one to complete an isoparaffin isomer. The constructed molecule is shipped to the MOPAC2002 software for geometry optimization and molecular pre-screening. As mentioned above, mono-methyl-branched isoparaffins are abundant in crudes but are not thermodynamically favored (see Figure 5.2). In case these molecules being screened out, they are excluded from pre-screening procedures and selected directly. The rest are screened using the same ΔH^0_f threshold. This procedure proceeds until the last isomer is constructed and simulated, as shown in the programming

flow chart in Figure 5.6. Then the program returns back to main program (Figure 5.5) for next #C isomeric lump until all the isoparaffins being constructed.

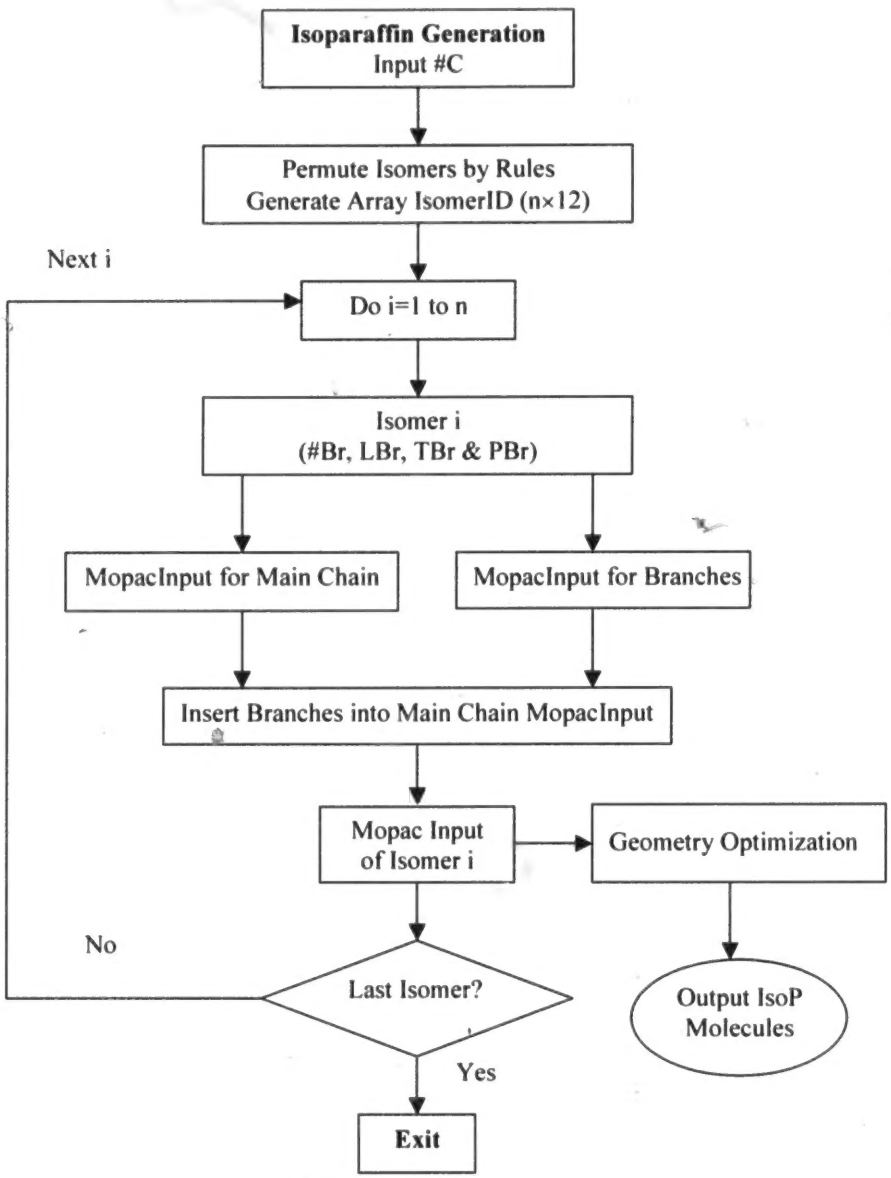
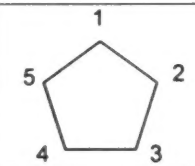
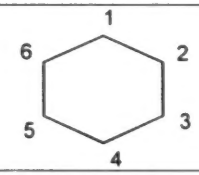
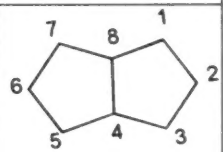
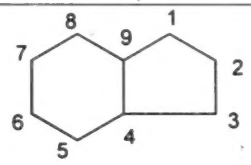
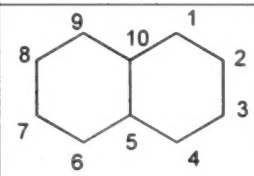
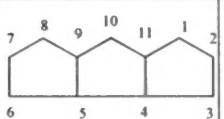
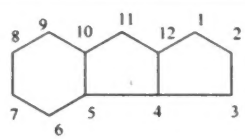
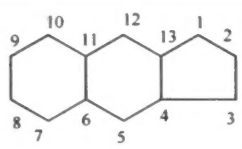
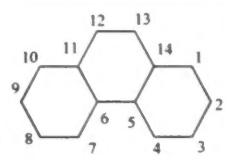
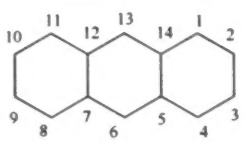
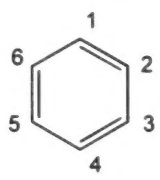
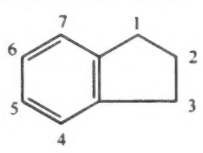
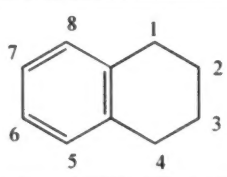
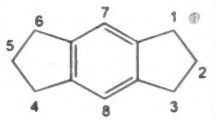
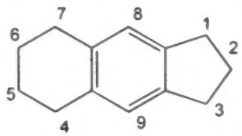
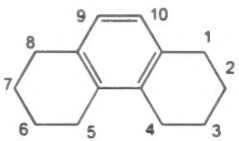
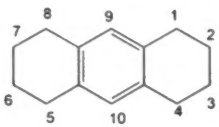
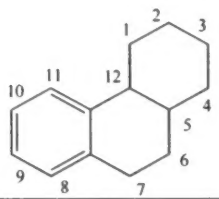
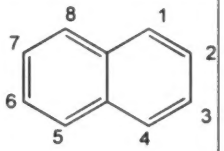
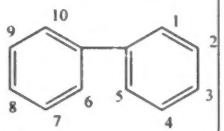
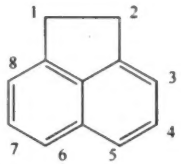
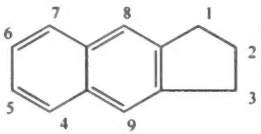
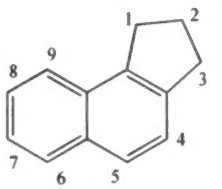
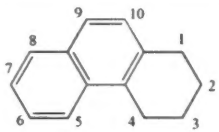
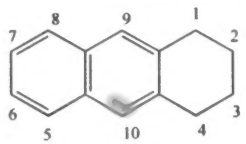
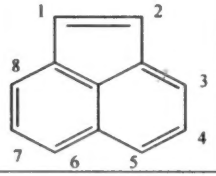
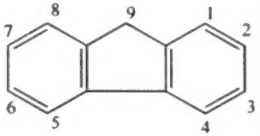
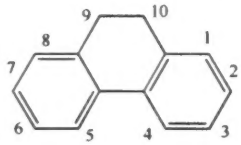
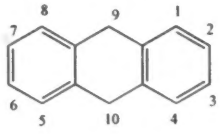


Figure 5-6. Programming diagram for iso-paraffin generation

5.2.2.4.2 Molecular generations of cyclic hydrocarbons

The principle for generating cyclic hydrocarbon molecules is similar to that for building iso-paraffins, except that the main structure here is a ring-core. Some cyclic hydrocarbons can have more than one type of ring-core structures at a given #C (here only the five-member or six-member ring is considered). For the cyclic hydrocarbons identified in GC-FIMS (Z=0 to –24) for diesels, the ring structure arrangements are limited even for tetra-aromatics (e.g., 5 ring-core structures for Z=-24). The (likely) ring-core structures, corresponding to the hydrocarbon types (HCType =3 to 16) determined by GC-FIMS for Z =0 to –24, are illustrated in Figure 5.7. For those homologous series having complex ring arrangements (Z=-4, -10, -14, -20, and –22), molecular simulations have been conducted for the possible ring arrangements and the most 5 thermodynamically favorable ones are selected as the ring-cores being utilized in the molecular assembling. As shown in Figure 5.8, generating cyclic hydrocarbons inputs HCType and #C. For a given HCType and #C, the permutation procedure will select a ring-core from Figure 5.7, accordingly, and then the possible isomers are permuted from rules for the selected ring-core. When the ring-core has just one methyl, one ethyl, or two methyl substitutes, the permutation goes through the non-symmetrically available positions and directly assign the a 2D matrix IsomerID (n×14) containing n isomers. The 14 entries of isomer i refer to IsomerID[i, (#C, #Br, 4LBrs, 4TBrs, 4PBrs)], respectively. If the total length of substitutes ≥ C3, the permutation selects the #Br =1, 2, 3/4 consequently, then permutes the branches over the available sites (as numbered on the ring-core in Figure 5.7).

Type	1	2	3	4	5
HCType =3 1N Z=0					
HCType =4 2N Z=-2					
HCType =5 3N Z=-4					
HCType =6 1A Z=-6					
HCType =7 1A1N Z=-8					

HCType =8 1A2N Z=-10					
HCType =9 2A Z=-12					
HCType =10 A_A Z=-14					
HCType =11 2A1N Z=-14					
HCType =12 ANA Z=-16					

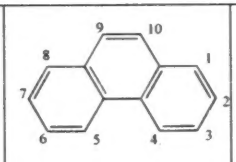
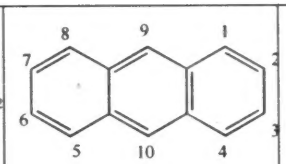
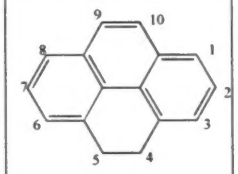
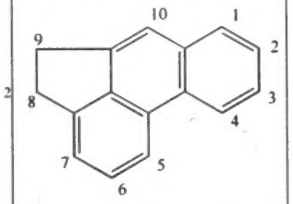
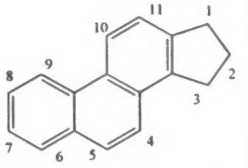
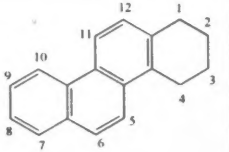
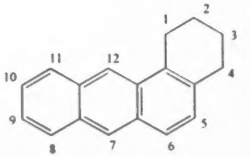
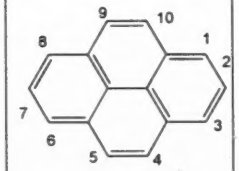
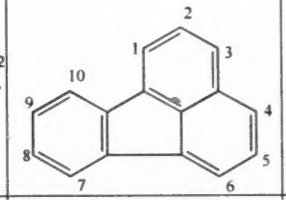
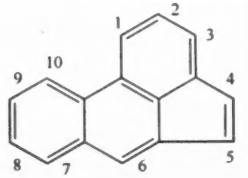
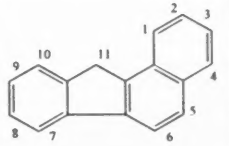
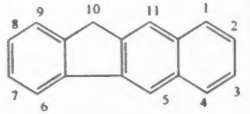
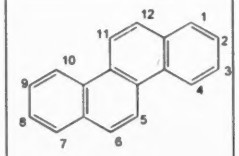
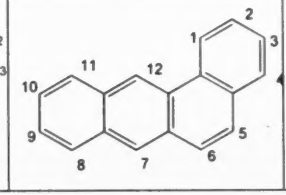
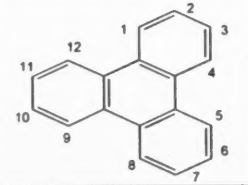
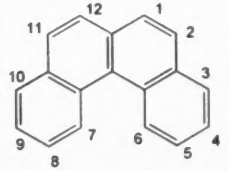
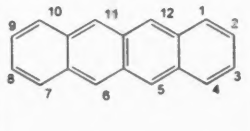
HCType =13 3A Z=-18					
HCType =14 3A1N Z=-20					
HCType =15 4Ap Z=-22					
HCType =16 4Ac Z=-24					

Figure 5-7. Ring-core structures of cyclic hydrocarbons representing molecules identified by GC-FIMS

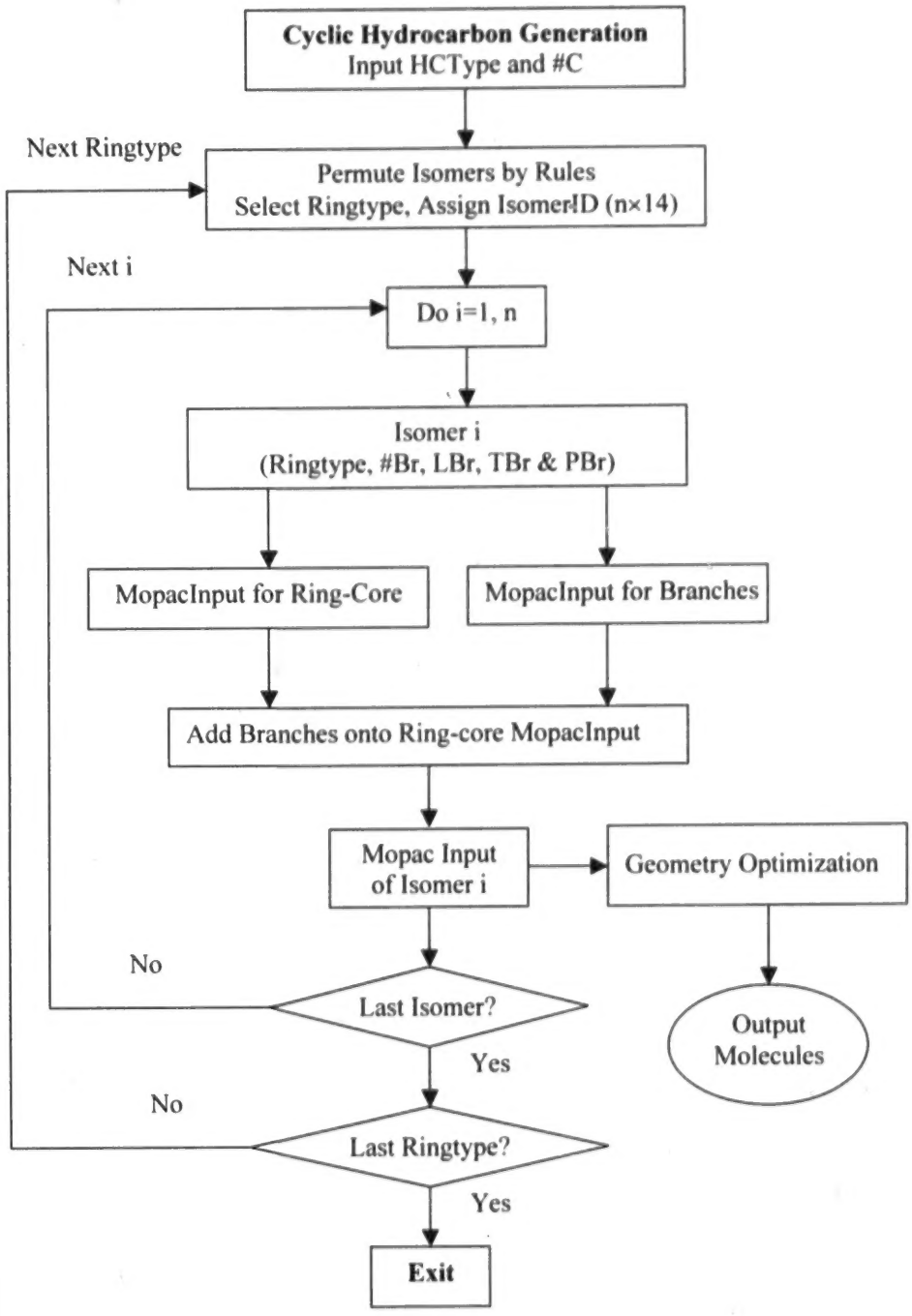


Figure 5-8. Programming diagram for generation of cyclic hydrocarbons

Following the molecular generation rules, the substitutes are distributed over the ring-core as apart as possible with sorted permutations of LBr, TBr, and PBr. The permuted results are assigned in the matrix IsomerID (i,j=1 to 14) with i accumulating from 1 to n isomers. The procedure proceeds until the last type/position being reached. The MOPAC input of the selected ring-core structure is generated first. Then, from the substitution attributes contained in IsomerID, cyclic hydrocarbons are generated by assembling the MOPAC input of the ring-core with the MOPAC input characteristics of the selected branches. The same types of substituents as being used for isoparaffins, were used for constructing cyclic hydrocarbons except that longer substituents (LBr ≥ 5) were allowed. Based on the occurrence of substituted cyclic structures and the purpose of simplification, only linear or iso-branched-linear substitutes of LBr ≥ 5 are utilized in the construction of cyclic molecules, as showed in Table 5.1.

The generated molecules are then shipped to "MOPAC2002" for on-line geometry optimization, and then molecular pre-screening. This assembling procedure proceeds until the last isomer is constructed. If there is more than one ring-core structures listed in Figure 5.7 for the selected HCType, the program goes to the next ring-core, repeats the permutation and assembling procedures described above until all the isomers for the last ring-core are generated. Then, the program returns to the main program (Figure 5.8) for next #C, next HCType, and so on. Once all the molecules, referring to the isomeric lumps identified in GC-FIMS report, are generated and selected after pre-screening, they are stored in the molecular bank in the format of MOPAC input with the updated 3D descriptions from PM3 optimization output. The physical properties (BP, density, and RI) and thermodynamic properties (entropy and enthalpy) are then calculated using developed QSPR correlations and MOPAC2002, respectively. With these properties

assigned, the selected molecules are ready to be manipulated for isomer distribution and bulk property predictions, as will be described shortly.

QSPR study of boiling point, specific gravity, and refractive index

After finding a molecular representation of a petroleum fraction, we are able to test its validity independently. In this work, this is achieved by comparing predicted and measured bulk physical properties of the fraction. When the molecular representation of the fraction is found, predictive models for *BP*, *SG*, and *RI* of pure compounds can be used to estimate bulk properties of that fraction using simple mixing rules. Unfortunately, most models for estimating these physical properties of pure hydrocarbons in the open literature do not offer the accuracy we pursue. Therefore, we developed a set of QSPR models for *BP*, *SG*, and *RI* estimation as introduced in Chapter 2 (Ha et al., 2005a). QSPR models have been successfully used in such areas as analytical chemistry and pharmaceutical research for estimating various properties of pure compounds. The QSPR approach relies on correlating the property in question with the molecular structure of the compound captured in the form of a number of descriptors. Those can be divided into several groups: constitution, topology, geometry, electrostatics, quantum-chemistry, and thermodynamic. Using statistical methods and large data sets including 186 saturated 200 aromatic hydrocarbons, we developed a set of empirical multi-linear models that were demonstrated to be significantly better than the group contribution models commonly used for property estimation. The quality of these models is reflected in Table 5.2. Except for the *SG* model for aromatics (R^2 = 0.9881), the correlation coefficients of the models were better than 0.99. The standard deviations of these models were less than ±6.2°C for

BP, ±0.008 for *SG*, and ±0.005 for *RI*, respectively. The leave-one-out cross validation and external test data sets were also used to validate the models with excellent results.

Table 5.2. QSPR models' performance on *BP*, *SG*, and *RI* predictions for hydrocarbons

Models	F#	R^2	$R^2(CV)^*$	*s*	*s*(test)	r%(test)+
BP for Sat.	13332.1	0.9979	0.9976	±5.87°C	±6.1°C	0.68
BP for Aro.	6351.6	0.996	0.9954	±6.2°C	±7.24°C	0.81
SG for Sat.	1925.4	0.991	0.9894	±0.007	±0.01	1.0
SG for Aro.	1820.9	0.9881	0.9863	±0.008	±0.01	0.8
RI for Sat.	3054.3	0.9921	0.9902	±0.004	±0.006	0.4
RI for Aro.	2052.26	0.9902	0.9882	±0.005	±0.01	0.53

#Fisher number, *Cross validation correlation coefficient, +Average relative error %

5.2.4 Isomer distribution within an isomeric lump

It is impractical to find detailed molecular composition of petroleum fractions boiling above 200°C experimentally. Modern analytical techniques provide at best compositional information in terms of concentrations of isomeric lumps. An intuitive way to "de-lump" this information is to distribute the concentration of the isomeric lump among individual isomers assuming thermodynamic equilibrium. Based on the assumption of ideal solution considered appropriate for petroleum mixtures, Smith and Missen (1982) presented a thermodynamically rigorous treatment of equilibrium among isomers in a closed system. However, in petroleum samples, the measured isomeric distributions are not consistent with thermodynamic equilibrium distributions. Compositional analyses of various crude oils reveal that the reported abundance of alkane isomers in a number of crude oils is quite far from the thermodynamic equilibrium distribution (Tissot and Welte, 1984). It was speculated that either the thermodynamic equilibrium has not reached or the isomers undergo other chemical reactions which violate the assumption of the closed system and

this makes the problem intractable because of the compositional complexity of petroleum (Alberty, 1991).

A molal average boiling point (BP_{lump}) was introduced to the fundamental chemical equilibrium algorithm, and helped us to match the experimental isomer distributions (Ha et al., 2005b). This BP_{lump} can be determined from GC-FIMS measurements with the aid of n-paraffin calibration standards. The BP_{lump} carries information about the composition of the isomeric lump and can be used as a constraint in the minimization of the free energy of a mixture of isomers. Then the problem becomes that of search for the equilibrium composition subject to the extra average boiling point constraint. Mathematically, this corresponds to constrained minimization of the free energy (ΔG_n^o) of the system

$$\min\left[\Delta G_n^0\right] = \min\left[\sum_{i=1}^{N} x_i \Delta G_i^0 + RT\sum_{i=1}^{N} x_i \ln x_i\right]$$
$$Subject\ \ to\ \ \begin{cases} \sum x_i = 1 \\ \sum x_i BP_i = BP_{lump} \end{cases} \qquad (5\text{-}2)$$

It was proven analytically that this problem has a unique solution (Ha et al., 2005b).

This approach was applied to predict the isomeric distribution among hexane/heptane isomers using standard free energy of formation and the normal boiling point of each isomer (Ha et al., 2005b). The predicted results were compared with the reported data of 18 geologically different crude oils (Martin et al., 1963). The calculated distributions of hexane/heptane isomers matched the reported data satisfactorily, except for three younger crudes formed during the tertiary Cenozoic Era (South Houston, wafra, and Wilminton). In particular, the approach can predict the key isomers with average relative prediction errors of 15% and 12% for hexane and heptane isomers, respectively. These key isomers,

3 out of 5 for hexanes and 4 out of 9 for heptanes, account for more than 92/90% of the total hexane/heptane isomers reported (Martin et al., 1963).

This approach is also used to de-lump the isomeric lump measured by GC-FIMS in this work. However, petroleum products (like gasoline or diesel) are usually blended from various processing streams, and it is impractical to trace back the formation conditions of the samples. It is difficult to determine the state of the mixture that dominates the distribution of isomers. As an alternative, the standard Gibbs free energy of formation is used in this work. On the other hand, as #C increases, single BP_{lump} may introduce the uncertainties in representing BP distribution due to the wide spread of BP and overlaps among isomers. These discrepancies introduced hereof are taken into account by an optimization procedure. Instead of using one averaged BP_{lump} (like the cases for hexane/heptane isomer distributions introduced in Chapter 3), this procedure takes a set of BP_{lump} within the boiling range of isomers and produces a set of solutions (**X**) for isomer distribution. The solutions are then compared with the measured isomer-BP-distribution (like SimDis) from GC-FIMS. The solution closest to the measured BP distribution is taken as the best distribution of isomers. As a result, the detailed molecular profile of petroleum mixtures could be accessed beyond analytical limits.

5.2.5 Derivation of minimum set molecular representation

With an initial value of molecular pre-screening threshold (2.0 kcal/mol in heat of formation used in this work), a molecular set can be selected as the initial molecular representation. Simple linear mixing rules work well in estimating bulk properties for petroleum fractions (Miquel and Castells, 1993). From this representation, bulk properties (SG, RI, and MW) of a mixture can be predicted using following mixing rules:

$$\frac{1}{SG} = \sum_i \frac{x_{wi}}{SG_i} \tag{5-3}$$

$$\frac{1}{MW} = \sum_i \frac{x_{wi}}{MW_i} \tag{5-4}$$

and

$$RI = \sum_i x_{wi} RI_i \tag{5-5}$$

where x_{wi} is the mass fraction of molecule *i*.

Sorting all the molecules by calculated BP and summing up the wt% in every 10°C cut will generate an equivalent SimDis curve. These bulk properties can be used for validation of current approach as compared with the experimental data. If the pre-screening threshold is appropriate, an adequate molecular representation can be obtained, from which good agreements can be found between the predicted bulk properties and the measured ones. Efforts were made to seek the minimum number of molecules without losing the consistency between the predicted and measured bulk properties. The less abundant molecules are eliminated by setting up a "doorstep" in concentration (e.g., 10^{-6} wt%). A tolerance (Tol) is defined to account the property changes due to the reduction of molecules by ignoring the concentrationally insignificant molecules:

$$Tol = |SG_i - SG_0| + |RI_i - RI_0| + |(Dev_SimDis)_i - (Dev_SimDis)_0| \tag{5-6}$$

and $$Dev_SimDis = \frac{1}{N}\sum_{j=1}^{N}\left(x_{w,pred} - x_{w,\exp}\right)_j \tag{5-7}$$

Where the subscripts *0* and *i* refer to the initial molecular set *0* and the minimized set *i*. The N is the number of cuts on experimental SimDis curve in every 10°C and x_w is the mass fraction of the cut *j*. The three terms in Equation 6 were tested to have equivalent

deviations corresponding to the reduction of molecules. Increasing the "doorstep" in concentration will eliminate more molecules until the preset tolerance is met. As a result, a minimum set of molecular representation is derived at a given tolerance.

5.3 Molecular representation and bulk property predictions

The molecular delumping algorithm shown in Figure 5.1 was applied to generate molecular representations of five diesel samples. The selected samples cover a variety of compositions (i.e. highly paraffinic - S1, highly naphthenic - S3, highly aromatic - S5). With the pre-screening threshold, ΔH_f^0 = 2.0 kcal/mol, initial sets of molecules, ranging from 801 to 1716 molecules, were generated and selected for sample 1 to 5, as listed in Table 5.3 respectively. Density and *RI* for sample 1 to 5 were calculated from their molecular representations using Equation 5-3 and 5-5. The average MW of sample 1 to 5 were directly estimated from their GC-FIMS characterization matrices like Table 4.4 using Equation 5-4 where x_{wi} is the measured mass concentration of isomeric lump *i*. These calculated properties agree well with the corresponding measurements for the five diesel samples, as compared in Table 5.3 as well. The maximum absolute prediction errors were ±0.008 g/ml for density and ±0.006 for RI, equivalent to the standard deviation in QSPR models for the pure compounds. The average prediction error for both bulk density and *RI* was only ±0.003 for all five samples. The agreement between the predicted and measured average Mw for sample 1 to 5, was found satisfactory with the average absolute error of ±2.44 amu comparable with the experimental error of the FPD method (1-2%). The maximum calculated deviation was only 4.9 amu.

Table 5.3 Simulated bulk properties compared with experimental results for 5 samples

Sample	Initial set	Mw, amu		ρ@15.6°C, g/ml		RI @25°C		Error in SimDis, °C		
		Meas.	Calc.	Meas.	Calc.	Meas.	Calc.	IBP	5-95%	FBP
S1	801	150	148.4	0.7985	0.7986	1.4436	1.4449	9.2	2.0	7.8
S2	1588	188	187.8	0.8291	0.8361	1.4593	1.4653	4.8	3.0	14.2
S3	1531	176	171.2	0.8357	0.8434	1.4584	1.4619	18.2	2.4	2.2
S4	1716	176	175.3	0.8476	0.8477	1.4652	1.4672	11.8	3.8	7.2
S5	1416	175	179.9	0.8537	0.8533	1.4755	1.4767	2.7	2.6	2.6

Furthermore, as another means of verifying the analytical data and the characterization method used in this work, two comparisons of the calculated and measured SimDis curves are shown in Figures 5-9 to 5-13. The first one is the comparison between the GC-FIMS generated (marked as "FIMS-gen SimDis") and the measured SimDis data, as introduced in Chapter 4. The second comparison is between the SimDis predicted from representative molecules (marked as "Simulation" in above Figures) and the measured ones. In general, good agreement was found. The deviations in IBP (0.5 wt% cut point) and FBP (99.5 wt% cut point) from simulations for each sample are also shown in Table 5.3, together with the average deviation over 5-95% recovery range. The average deviations for IBP and FBP over all five samples were 9.3°C and 6.8°C, respectively, where the average deviation in the 5-95% recovery range was only 2.7°C. For comparison, the repeatabilities of ASTM D2887 SimDis measurements were 6.0°C and 5.0°C for IBP and FBP, respectively. The repeatability in the 5-95% recovery range was 2.5°C. The corresponding reproducibilities were 23.0°C and 13.5°C for IBP and FBP, and 6.5°C over the 5-95% range. Consequently, the agreement for simulated results was found excellent. The good agreement between predicted and measured bulk properties indicates that ΔH_f^0 = 2.0 kcal/mol is an appropriate threshold for pre-screen hydrocarbon molecules and it allows adequate molecules to predict the bulk properties of diesels accurately.

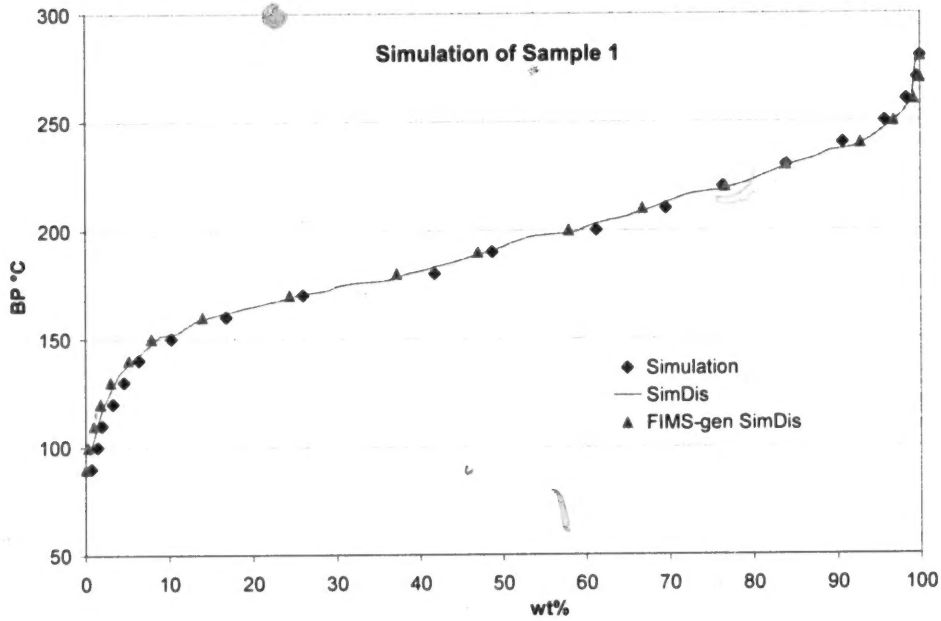


Figure 5.9 GC-FIMS-generated and Simulated SimDis compared with SimDis for sample 1

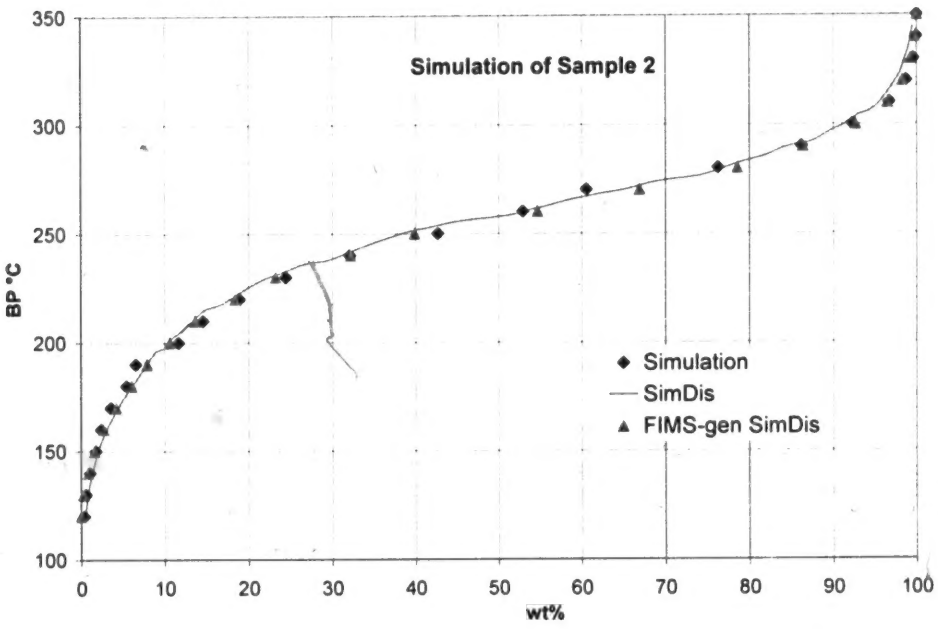


Figure 5.10 GC-FIMS-generated & Simulated SimDis compared with SimDis for sample 2

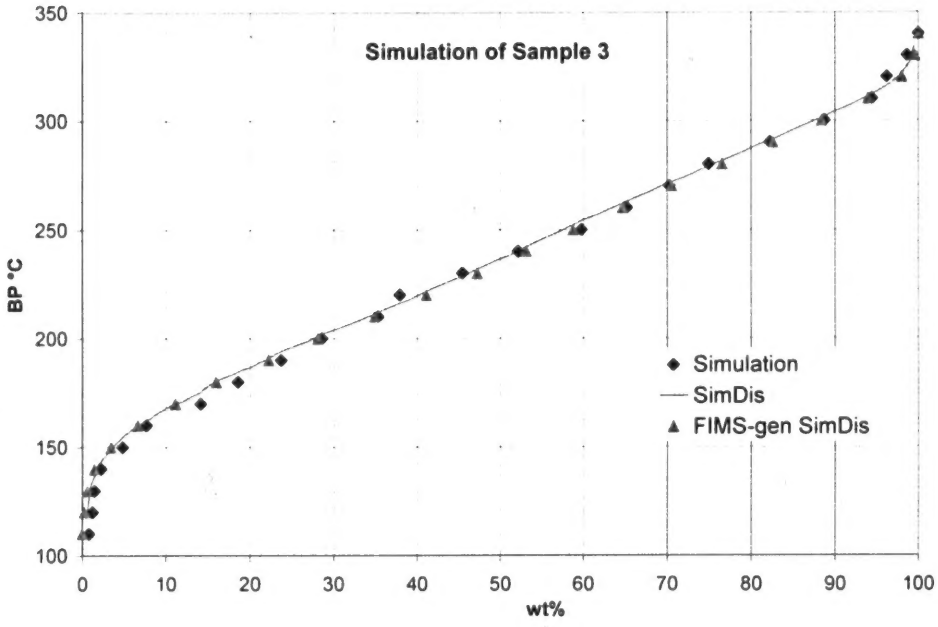


Figure 5.11 GC-FIMS-generated & Simulated SimDis compared with SimDis for sample 3

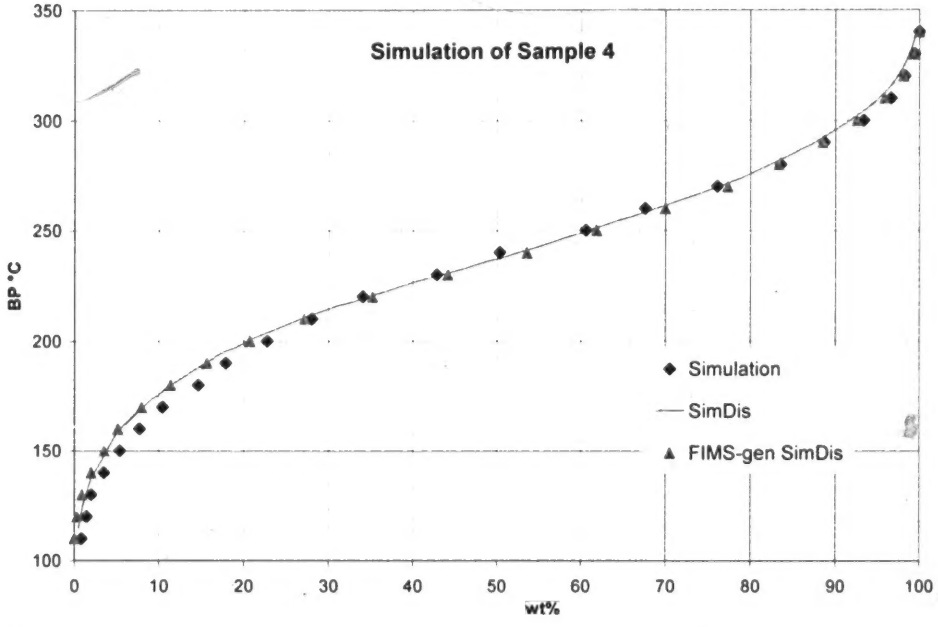


Figure 5.12 GC-FIMS-generated & Simulated SimDis compared with SimDis for sample 4

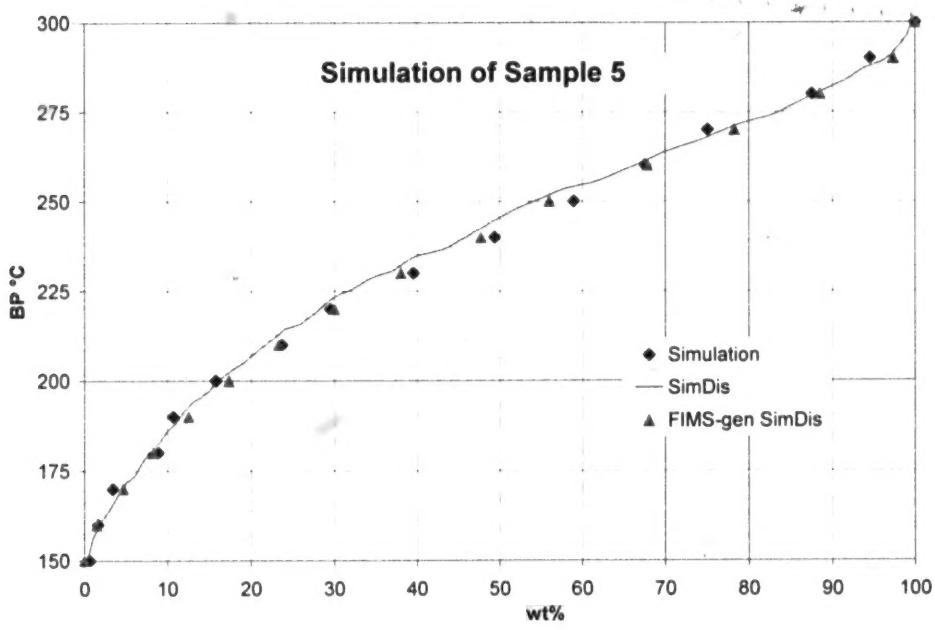


Figure 5.13 GC-FIMS-generated & Simulated SimDis compared with SimDis for sample 5

5.4 Sensitivity of Estimated properties to the size of molecular representation

We explored the consequences of reductions in the number of molecules in the representation for the estimated bulk properties. In this work, a tolerance of 10^{-4} was used and the number of molecules was reduced by 12% (sample 1) to 37% (sample 3) as indicated in Table 5.3 and 5.4. Table 5.4 lists the minimum molecular set of each molecular class (homologous series), given a tolerance of 10^{-4}. An overall mass balance was checked on the minimized molecular set. The mass losses of the eliminated molecules were insignificant (less than 0.2 wt%) for all five samples. Increasing a "doorstep" in concentration is equivalent to tightening the ΔH_f^0 threshold in molecular selection. Therefore, instead of regenerating and reselecting the molecules by changing ΔH_f^0 threshold, which is computationally demanding, we directly ignore those molecules

concentrationally insignificant to the bulk properties and achieve the minimum size of molecular representations. Since all the representative molecules are stored in MOPAC input format, their structures can be easily viewed using commercial software like Chem3D or Winpac. These selected molecules with calculated concentration and properties, so far, present a detailed molecular interpretation of petroleum mixtures with a high accuracy. Further application of these molecules can be extended to the detailed mechanistic kinetic models and the refinery process separation models.

Table 5.4 Simulated results with minimum set of representative molecules

Hydrocarbon Class/Sample	S1	S2	S3	S4	S5
n-paraffin	10	15	13	12	8
Iso-paraffin	161	270	102	164	215
monocycloalkane	141	223	202	201	169
Dicycloalkane	124	199	182	193	158
Tricycloalkane	41	126	99	113	75
Benzene	111	182	150	159	141
BenzoCycloalkane	83	144	87	141	112
BenzoDiCycloalkane	12	76	52	65	31
Naphthalene	15	67	45	36	32
Biphenyl	6	53	20	19	23
Naphthocycloalkane	1	23	10	8	2
Fluorene	0	19	5	11	1
Phenanthrene	0	5	0	0	0
Total Number of Molecules	705	1402	967	1122	967

5.5 Conclusions

Based on the molecular generation rules, molecules have been constructed in MOPAC input format, simulated on-line, selected according to their thermodynamic stabilities, assigned concentration assuming quasi-equilibrium distribution among isomers, and optimized with the minimum number of molecules to represent the molecular compositions of petroleum mixtures. The detailed 3D profiles in MOPAC input present four new features in molecular characterizations: (1) presenting a detailed 3D description of molecular structures, (2) enabling of molecules to be viewable using

Chem3D or other commercial software; (3) making the on-line molecular simulation possible; (4) allowing QSPR models to be developed directly and to be used on-line.

All molecules have been geometrically optimized and selected based on molecular simulation, neither of which has been done in any published reports for molecular characterization. Therefore, the molecules presented in this work are most likely to exist in petroleum oils according to their reported occurrences and thermodynamic stabilities. Compared to the arbitrarily selected or randomly constructed molecules reported earlier, the current molecular generation and selection scenario gives an insight into the real compositional make-up of petroleum mixtures. The final set of representative molecules provides the most detailed and accurate compositional profile so far for petroleum fractions up to middle distillates. In the bulk-property-based random characterization methods, the same bulk properties were used for both development and validation of the molecular representations. Contrast to those works, the property estimation is an independent validation of the algorithm in this work, it could be used to test the assumptions made in the development of the molecular representation method.

In principle, for the heavy materials like vacuum residue and asphaltene, for which only those bulk properties (e.g., density, MW, elemental and NMR analyses) are available from analytical measurements, Monte Carlo simulation probably presents the best way to derive the detailed molecular representations (Neurock et al., 1994). However, whenever the detailed compositional mass-matrix (like GC-FIMS report in this work) is available from analytical measurements, such extensive and valuable information should be used in developing the molecular representations. The current work provides an effective way to do this. With the recent advances in analytical chemistry (Liang and Hsu, 1998; Wu et al., 2004), such detailed compositional mass-

matrix will be available for heavy material such as vacuum gas oil and heavy crude oils in the near future.

Using the current delumping algorithm, the valuable information from GC-FIMS by #C report is conserved. The algorithm has been applied to five diesel samples with different compositions. The final set of molecules not only match the measured bulk density, RI, and SimDis, but also match the hydrocarbon class contents and their distribution by #C and by BP for all the samples, which cannot be achieved by the bulk property-based characterization methods (e.g., using elemental and NMR data only). The proposed approach was proven valid for the diesel samples tested. However, with the available GC-FIMS data for a higher BP range, the current algorithm can be extended to the heavier samples like vacuum gas oil, adding a few more homologous series into the molecular bank (e.g., tetra-aromatics and aromatic sulfur compounds).

With the detailed 3D structural information and the calculated thermodynamic properties (e.g., heat of formation, Gibbs free energy of formation) for each molecule, the current molecular characterization algorithm can be readily applied to the detailed mechanistic kinetic models, like the LFER model and the Single-Event kinetic model. With the high accuracy of the predicted BP, the molecules can be lumped into Pseudo-molecules, in terms of BP cuts, which then can be used in commercial refinery processing software like Hysys/Aspen for flash calculations. Integration of the kinetic model with the separation model will lead to process optimization, where detailed molecular characterization like this work is essential. Other properties should be introduced in the future. Those for diesel range would include Cetane Number or Cloud Point. Of course, it would also be necessary to develop the QSPR correlations as needed.

5.6 Reference

Alberty, R. A. Chemical equilibrium in complex organic systems with carious choices of independent variables. *Chemical Reactions in Complex Mixtures*, edited by Sapre, A. V.; Krambeck, F. J. Van Nostrand Reinhold, New York, 1991.

API technical data Book, American Petroleum Institute, Washington, D.C, 1992.

Annual Book of ASTM Standards, Vol. 5, ASTM, PA, 2001.

Boduszynski, M. M. (1988). Composition of heavy petroleums. 2. molecular characterization. *Energy & Fuels*, **2**(5), 597-613.

Bouquet, M.; Brument, J. (1990). Characterization of heavy hydrocarbon cuts by mass spectrometry. Routine and quantitative measurements. *Fuel Sci. Tech. Int.* **8**(9), 961-986.

Briker, Y.; Ring, Z.; Iacchelli, A.; McLean, N.; Rahimi, M.; Fairbridge, C. (2001). Diesel fuel analysis by GC-FIMS: aromatics, n-paraffins, and isoparaffins. *Energy and Fuel*, **15**, 23-37.

Broadbelt, L. J.; Stark, S. M.; Klein, M. T. (1994a). Computer generated pyrolysis modeling: On-the-fly generation of species, reactions, and rates, *Ind. Eng. Chem. Res.*, **33**, 790-799.

Broadbelt, L. J.; Stark, S. M.; Klein, M. T. (1994b). Computer generated reaction network: On-the-fly calculation of species properties using computational quantum chemistry, *Chem, Eng. Sci.*, **49**, 4991-5010.

Campbell, D. M.; Klein, M. T. (1997). Construction of a molecular representation of a complex feedstock by Monte Carlo and quadrature methods, *Applied Catalysis A: General*, **160**, 41-54.

Chasey, K. L.; Aczel, T. (1991). Polycyclic aromatic structure distributions by high-resolution mass spectrometry, *Energy & Fuels*, **5**, 386-394.

Ha, Z.; Ring, Z.; Liu, S. (2005a). Quantitative Structure-property Relationship (QSPR) models for boiling points, specific gravities, and refractive indices of hydrocarbons, *Energy & Fuels*, **19**, 152-163.

.Ha, Z.; Ring, Z.; Liu, S. (2005b). Estimations of hydrocarbon isomer distribution in petroleum mixtures, *Energy and Fuels*, **19**(4), accepted.

Hillewaert, L. P.; Dierickx, J. L.; Froment, G. F. (1988). Computer generation of reaction scheme and rate equations for thermal cracking, *AIChE J.* **34**, 17-24.

Hood, A.; Clere, R. J.; O'Neal, M. J. (1959). The molecular structure of heavy petroleum compounds. *J. Inst. Petrol.*, **45** (426), 168-173.

Japanwala, S; Chung, K. H.; Dettman, H. D.; Gray, M. R. (2002). Quality of distillates from repeated recycle of residue, *Energy and Fuels*, **16**, 477-484.

Khan, H. U.; Sharma, R. L.; Nautiyal, S. P.; Agrawal, K. M.; Schmidt, P. (2000). Structural characterization of solid n-paraffin components derived from Mukta crude oil, *Petroleum Science and Technology*, **18**, 889-899.

Khorasheh, F; Khaledi, R.; Gray, M. R. (1998). Computer generation of representative molecules for heavy hydrocarbon mixtures, *Fuel*, **77**, 247-253.

Kim, H. N.; Verstraete, J. P.; Virk, P. S.; Fafet, A. (1998). NMR enhances mass-spec FCC feedstocks characterization. *Oil & Gas Journal*, **96**, 85-88.

Liang, Z.; Hsu, C.S. (1998). Molecular speciation of saturates by on-line liquid charomatography-field ionization mass spectrometry, *Energy & Fuels*, **12**, 637-643.

Mair, B. J.; Barnewall, J. M. (1964). Composition of the monoaromatic material in the light gas oil range, low refractive index portion, 230 to 305°C. *J. of Chem. Eng. Data*, **9**(2), 282-292.

Martin, R. L.; Winters, J. C.; Williams, J. A. (1963). Composition of crude oils by gas chromatography: geological significance of hydrocarbon distribution. 6th World Pet. Congr. Sec. **V**, 231-260.

Miquel, J.; Castells, F. (1993). Easy characterization of petroleum fractions, *Hydrocarbon Processing*, **Dec**., 101-105.

Mizan, T. I.; Klein, M. T. (1999). Computer-assisted mechanistic modeling of n-hexadecane hydroisomerization over various bifunctional catalysts. *Catalysts Today*, **50**, 159-172.

Neurock, M. N.; Nigam, A.; Trauth, D.; Klein, M. T. (1994). Molecular representation of complex hydrocarbon feedstocks through efficient characterization and stochastic algorithms, *Chem. Eng. Sci.*, **49**, 4153-4177.

Quann, R. J; Jaffe, S. B. (1992). Structure-oriented lumping: Describing the chemistry of complex hydrocarbon mixtures, *Ind. Eng. Chem. Res.*, **31**, 2483-2497.

Read, R. C. The enumeration of acyclic chemical compounds, *Chemical Application of Graph Theory*, edited by Balaban, A. T., Academic Press Inc., 1976.

Ramaswamy, V.; Singh, I. D.; Krishna, R. (1989). Characterization of vacuum gas oil from North Gujarat crude mix, *Indian Journal of Technology*, **27**, 85-88.

Smith, W. R.; Missen, R. W. *Chemical reaction equilibrium analysis: theory and algorithm*. John Wiley & Sons, New York, 1982.

Sheremata, J. M.; Gray, M. R.; Dettman, H. D.; and McCaffrey, W. C. (2004). Quantitative Molecular Representation and Sequential Optimization of Athabasca Asphaltenes, *Energy & Fuels*, **18**, 1377-1384.

Speight, J. G. *The chemistry and Technology of Petroleum*, Marcel Dekker, Inc., New York, 3rd ed., 1999.

Stewart, J. J. P. Semiempirical molecular orbital methods, in *Reviews in Computational Chemistry* (Edited by D.B. Lipkowitz and D. B. Boyd), Chapter 2. VCH Publishers, New York, 1990.

Stull, D. R.; Westrum Jr., E. F.; Sinke, G. C. *The chemical thermodynamics of organic compounds*, John Wiley & Sons, Inc. New York, p205, 1969.

Tissot, B. P.; Welte, D. H. *Petroleum Formation and Occurrence*, Springer-Verlag, 1984.

Wade JR., L. G. *Organic Chemistry*, p132, Prentice-Hall, New Jersey, 1987.

Wojciechowski, B. W. (1998). The reaction mechanism of catalytic cracking: quantifying activity, selectivity, and catalyst decay. *Catal. Rev.-Sci. Eng.*, **40**(3), 209-328.

Wu, Z.; Rodgers, R. P.; Marshall, A. G. (2004). Two- and three-dimensional van Krevelen diagrams: a graphical analysis complementary to the Kendrick mass plot for sorting elemental compositions of complex organic mixtures based on ultrahigh-resolution broadband Fourier Transform Ion Cyclotron Resonance mass measurements, *Anal. Chem.*, **76**, 2511-2516.

Chapter 6.

Conclusions and Recommendations

6.1 Conclusions

A unique deterministic method has been presented for molecular characterization of petroleum mixtures. The method is an effective approach to derive molecular representations when the detailed compositional mass-matrix information is available. The proposed characterization method directly utilizes the GC-FIMS report – the hydrocarbon type distribution by #C – that represents the current analytical limit. In this research, molecular generation rules are stipulated from the reported occurrence and molecular simulations. The hydrocarbon molecules are constructed from the generation rules, and simulated on-line by MOPAC2002. Then the thermodynamically favored molecules are selected to represent the constituents of the petroleum mixture. The number of molecules is controlled by an adjustable threshold in terms of difference of heat of formations. This criterion is an effective way for selecting and screening molecules.

The major contribution of this work is the finding of an isomer distribution method that enables to delump the isomeric lumps measured by GC-FIMS. Within an isomeric lump, the isomer distribution can be determined by minimizing the free energy of formation of the lump, subjected to the stoichiometric constraint and a measurable average BP_{lump} constraint. This approach has been used to predict the heptane isomer distribution in crudes and has been compared with the reported data. A good agreement was found between the predicted and the reported distributions over 18 crude oils. This approach has also been applied to five diesel samples, and incorporated with the

molecular generation and selection. The predicted bulk properties are consistent with the experimental data for all five diesels. These bulk properties are not involved in the characterization procedures and serve as independent validation checks. A good consistency has also been found between the calculated SimDis and the measured one for each sample. This further validates the proposed isomer distribution approach, and thus, the characterization method.

A special feature in this molecular representation is that, for the first time, molecules are generated in MOPAC input format and on-line molecular simulations are associated with generated molecules. The MOPAC input format provides detailed 3D descriptions of the molecular structure and, thus, allows an on-line geometry optimization by MOPAC or other quantum chemistry software. The simulated results (MOPAC output), in turn, are directly used for molecular selection and QSPR predictions. As a result, an automation process is created for overall molecular characterization. As a bonus of this special feature, the MOPAC input file can be viewed by software like ChemOffice or CODESSA. This makes the generated molecules more tangible and user-friendly.

There are correlations among various measured properties involving different instruments and analytic methods. For example, the GC-FIMS or GC-MS results can be converted to an equivalent SimDis curve. This study has also employed, for the first time in molecular characterization, checks on the consistency among the analytic data used. The data reconciliation approach in this thesis provides a self-consistency check on BP-RT calibrations and a cross-consistency check between the measured SimDis and the GC-FIMS generated SimDis. These consistency checks serve as a QC step in the characterization procedure, and ensure reliable supporting data for molecular characterization.

In facilitating the predictions of the bulk properties of petroleum mixtures, the physical properties of pure hydrocarbons have been studied using Quantitative Structure-Property Relationships. Multi-linear correlations for BP, density, and refractive index (RI) have been developed to predict these properties for each molecule generated. These QSPR models are more accurate than the traditional group contribution methods. Large training data sets have been used for model development and separated test sets were used to check the model predictive abilities. The relative prediction errors are less than 1% for all the models. These QSPR models not only provide an effective way for estimating the physical properties of hydrocarbons, but also elucidate the relationship between the molecular structure and specific properties. The geometry-optimized molecules are shipped into these QSPR models (in MOPAC output format) automatically via CODESSA and on-line predictions are made.

Generally, with the aid of this computer technique, an automated molecular characterization method has been developed for petroleum middle distillates. The approach takes the reconciled GC-FIMS report as input, automatically generates, simulates, and selects the molecules, and then estimates the physical properties, distributes the isomers within the measured isomeric lump, calculates the bulk properties and SimDis, and finally compares with the measured ones to validate the method. It offers an accurate compositional profile for petroleum mixtures at the molecular level and has potential applications in mechanistic kinetic modeling and separation calculations.

6.2 Recommendations

Due to the current availability of GC-FIMS data at NCUT, the developed characterization approach has only been applied to the middle distillates (Diesels). NCUT

is now installing a new GC-FIMS that will extend the handling of material up to the Vacuum Gas Oil range. Once the new GC-FIMS data is available for heavier material like VGO, the current approach can be extended to characterize these heavy materials. To facilitate characterizing VGO, several hydrocarbon types (homologous series) need to be added into the molecular generation matrix, such as penta-aromatics and aromatic sulfur compounds. In diesels, these compounds are either not present or their amounts are negligible.

So far, this characterization method has only been used to predict the several bulk properties (density and RI) of mixtures, and the results are good. More valuable properties like Octane Number for gasoline and Cetane Number for diesels can be estimated if sufficient data can be found for developing a QSPR model to predict these properties for a pure compound. The motivation of this work was to develop a detailed molecular representation that can be implemented into the detailed mechanistic kinetic models for hydrotreating, hydrocracking, or catalytic cracking processes. With the MOPAC input format and the thermodynamic properties associated with each representative molecule, the current molecular characterization can be easily applied to such complex models once they are available. Further application can be extended to overall process optimization to improve refinery profits.

Appendix A. The raw data set for QSPR correlations

Table A1. The 481-hydrocarbon data set with reported and predicted *BP*, *SG*, and *RI*

Satuarate Training Set (N=186)

Structure	Cal. BP(K)	Exp. BP(K)	Cal. SG	Exp. SG	Cal. RI	Exp. RI	Ref.
np-C1	116.09	111.66	0.2998	0.2999	0.9966	1.0004	a
np-C2	179.81	184.55	0.3651	0.3554	1.1966	1.1849	a
np-C3	219.58	231.11	0.4976	0.5063	1.2773	1.2861	a
np-C4	270.06	272.65	0.5704	0.5849	1.3215	1.3292	a
np-C5	307.72	309.22	0.6180	0.6317	1.3462	1.3547	a
np-C6	346.18	341.88	0.6540	0.6651	1.3691	1.3723	a
np-C7	374.31	371.58	0.6807	0.6902	1.3809	1.3851	a
np-C8	402.85	398.83	0.7005	0.7073	1.3929	1.3951	a
np-C9	427.41	423.97	0.7176	0.7220	1.4011	1.4031	a
np-C10	450.22	447.31	0.7302	0.7342	1.4083	1.4097	a
np-C11	471.12	469.08	0.7419	0.7439	1.4142	1.4151	a
np-C12	491.05	489.47	0.7508	0.7524	1.4194	1.4151	a
np-C13	509.29	508.62	0.7590	0.7611	1.4235	1.4235	a
np-C14	526.60	526.73	0.7655	0.7665	1.4273	1.4268	a
np-C15	542.86	543.84	0.7718	0.7717	1.4303	1.4298	a
np-C16	558.39	560.01	0.7765	0.7760	1.4331	1.4325	a
np-C17	573.67	575.30	0.7815	0.7753	1.4355	1.4348	a
np-C18	587.07	589.86	0.7850	0.7841	1.4376	1.4369	a
np-C19	601.41	603.05	0.7887	0.7880	1.4393	1.4388	a
np-C20	614.76	616.93	0.7919	0.7890	1.4412	1.4405	a
np-C21	626.80	629.65	0.7947	0.7954	1.4424	1.4420	a
np-C22	639.10	641.75	0.7972	0.7981	1.4439	1.4434	a
np-C23	650.42	653.35	0.7992	0.8123	1.4447	1.4447	a
np-C24	661.85	664.45	0.8016	0.8027	1.4459	1.4459	a
np-C25	672.49	675.05	0.8031	0.8048	1.4467	1.4470	a
np-C26	683.07	685.35	0.8051	0.8067	1.4475	1.4481	a
np-C27	693.34	695.25	0.8063	0.8085	1.4483	1.4491	a
np-C28	703.25	704.75	0.8081	0.8077	1.4488	1.4499	a
np-C29	713.31	713.95	0.8088	0.8120	1.4498	1.4509	a
np-C30	722.59	722.85	0.8106	0.8123	1.4500	1.4515	a
Isobutane	264.24	261.43	0.5708	0.5644	1.3232	1.3175	a
Isopentane	295.87	300.99	0.6227	0.6265	1.3482	1.3509	a
Neopentane	299.91	285.65	0.6271	0.6073	1.3531	1.3390	a
2M-pentane	331.01	333.41	0.6551	0.6577	1.3665	1.3687	a
3M-PENTANE	330.51	336.42	0.6652	0.6693	1.3726	1.3739	a
2,2MM-BUTANE	331.03	322.88	0.6702	0.6539	1.3755	1.3659	a
2,3MM-BUTANE	327.21	331.13	0.6613	0.6662	1.3686	1.3728	a
2M-hexane	361.07	363.20	0.6815	0.6822	1.3795	1.3823	a
3M-hexane	360.79	365.00	0.6883	0.6922	1.3839	1.3861	a
3E-pentane	359.44	366.62	0.6996	0.7043	1.3883	1.3908	a
2,2MM-pentane	356.17	352.34	0.6821	0.6818	1.3799	1.3795	a

2,3MM-pentane	359.79	362.93	0.6996	0.6994	1.3898	1.3895	a
2,4MM-pentane	357.57	353.64	0.6779	0.6764	1.3772	1.3788	a
3,3MM-pentane	355.93	359.21	0.7028	0.6961	1.3927	1.3884	a
2,2,3MMM-Butane	358.64	354.03	0.7041	0.6954	1.3925	1.3869	a
2M-HEPTANE	391.25	390.80	0.7036	0.7029	1.3923	1.3926	a
3M-HEPTANE	389.53	392.08	0.7076	0.7092	1.3941	1.3961	a
4M-HEPTANE	389.29	390.86	0.7102	0.7096	1.3949	1.3955	a
3E-HEXANE	387.43	391.69	0.7133	0.7173	1.3975	1.3992	a
2,2MM-HEXANE	379.84	379.99	0.6988	0.7002	1.3878	1.3910	a
2,3MM-HEXANE	383.89	388.76	0.7140	0.7162	1.3973	1.3988	a
2,4MM-HEXANE	384.36	382.58	0.7033	0.7017	1.3916	1.3929	a
2,5MM-HEXANE	381.91	382.26	0.7050	0.6983	1.3921	1.3900	a
3,3MM-HEXANE	382.62	385.12	0.7096	0.7141	1.3966	1.3978	a
3,4MM-HEXANE	384.59	390.88	0.7263	0.7243	1.4043	1.4018	a
2M-3E-Pentane	385.96	388.80	0.7207	0.7240	1.3997	1.4017	a
3M-3E-Pentane	386.49	391.42	0.7308	0.7317	1.4087	1.4055	a
2,2,3MMM-PENTANE	383.05	383.00	0.7222	0.7200	1.4008	1.4007	a
2,2,4MMM-PENTANE	378.43	372.39	0.7003	0.6988	1.3876	1.3890	a
2,3,3MMM-PENTANE	380.22	387.92	0.7289	0.7301	1.4030	1.4052	a
2,3,4MMM-PENTANE	384.62	386.62	0.7193	0.7240	1.3982	1.4020	a
2M-OCTANE	415.91	416.43	0.7177	0.7176	1.3998	1.4008	a
3M-OCTANE	416.11	417.38	0.7257	0.7247	1.4038	1.4040	a
4M-OCTANE	415.19	415.59	0.7238	0.7243	1.4036	1.4039	a
3E-HEPTANE	415.54	416.35	0.7304	0.7303	1.4072	1.4070	a
2,2MM-HEPTANE	406.40	405.84	0.7136	0.7146	1.3971	1.3993	a
2,6MM-HEPTANE	409.79	408.36	0.7191	0.7137	1.4008	1.3983	a
2,2,3MMM-HEXANE	408.72	406.73	0.7367	0.7336	1.4087	1.4082	a
2,2,4MMM-HEXANE	407.47	399.69	0.7276	0.7197	1.4008	1.4010	a
2,2,5MMM-HEXANE	401.90	397.24	0.7175	0.7119	1.3975	1.3973	a
2,3,3MMM-HEXANE	408.53	410.83	0.7408	0.7419	1.4130	1.4119	a
2,3,5MMM-HEXANE	404.23	404.51	0.7322	0.7261	1.4036	1.4037	a
2,4,4MMM-HEXANE	406.47	403.81	0.7303	0.7281	1.4051	1.4052	a
3,3,4MMM-HEXANE	407.13	422.60	0.7516	0.7498	1.4165	1.4154	a
3,3EE-PENTANE	408.67	419.34	0.7457	0.7575	1.4155	1.4184	a
2,2MM-3E-PENTANE	406.91	406.99	0.7388	0.7390	1.4077	1.4010	a
2,4MM-3E-PENTANE	410.63	409.87	0.7443	0.7423	1.4112	1.4115	a
2,2,3,3MMMM-PENTANE	411.95	413.44	0.7570	0.7607	1.4218	1.4214	a
2,2,3,4MMMM-PENTANE	407.68	406.18	0.7396	0.7430	1.4090	1.4125	a
2,2,4,4MMMM-PENTANE	406.90	395.44	0.7207	0.7236	1.4004	1.4046	a
2,3,3,4MMMM-PENTANE	406.94	414.70	0.7606	0.7588	1.4212	1.4200	a
2M-NONANE	440.58	440.15	0.7332	0.7307	1.4081	1.4075	a
3M-NONANE	438.44	440.95	0.7388	0.7369	1.4097	1.4103	a
4M-NONANE	437.68	438.85	0.7361	0.7361	1.4090	1.4095	a
5M-NONANE	438.23	438.30	0.7364	0.7363	1.4096	1.4100	a
2,7MM-OCTANE	431.75	433.02	0.7341	0.7279	1.4077	1.4062	a

3,3,4MMM-HEPTANE	431.32	435.05	0.7581	0.7607	1.4203	1.4213	a
3,3,5MMM-HEPTANE	431.56	428.83	0.7508	0.7469	1.4165	1.4147	a
2,2,3,3MMMM-HEXANE	427.74	433.46	0.7561	0.7684	1.4198	1.4260	a
2,2,5,5MMMM-HEXANE	426.10	410.61	0.7329	0.7229	1.4059	1.4032	a
2,4MM-3IsoP-PENTANE	430.97	430.19	0.7571	0.7624	1.4171	1.4225	a
2,2MM-OCTANE	428.74	430.05	0.7294	0.7285	1.4038	1.4060	a
"2M-decane"	462.96	451.05	0.7437	0.7520	1.4141	1.4201	e
"2M-undecane"	482.39	478.15	0.7534	0.7494	1.4188	1.4232	e
"2M-dodecane"	502.00	499.15	0.7616	0.7568	1.4232	1.4241	e
"2M-tridecane"	519.44	518.15	0.7679	0.7634	1.4268	1.4259	e
"2M-tetradecane"	535.91	534.15	0.7732	0.7693	1.4292	1.4284	e
"2M-pentadecane"	552.65	553.95	0.7793	0.7741	1.4329	1.4310	e
"2M-hexadecane"	568.07	568.15	0.7827	0.7793	1.4352	1.4336	e
"2M-heptadecane"	581.47	584.15	0.7870	0.7838	1.4367	1.4386	a
CycloPenatne	317.54	322.40	0.7459	0.7502	1.4060	1.4036	a
M-CycloPentane	346.76	344.96	0.7516	0.7540	1.4029	1.4070	a
E-CyCloPentane	376.55	376.62	0.7820	0.7712	1.4211	1.4173	a
1,1MM-Cyclopentane	356.48	361.00	0.7673	0.7593	1.4178	1.4109	a
cis-1,2MM-Cyclopentane	367.25	372.68	0.7744	0.7771	1.4178	1.4196	a
trans-1,2MM-Cyclopentane	365.39	365.02	0.7672	0.7561	1.4142	1.4094	a
cis-1,3MM-Cyclopentane	366.83	363.92	0.7660	0.7496	1.4155	1.4063	a
trans-1,3MM-Cyclopentane	363.24	364.88	0.7665	0.7534	1.4140	1.4081	a
Propyl-Cyclopentane	409.22	404.11	0.7756	0.7811	1.4201	1.4239	a
IsoP-Cyclopentane	394.61	399.58	0.7771	0.7806	1.4226	1.4235	a
1M-1E-Cyclopentane	395.54	394.67	0.7788	0.7853	1.4235	1.4248	a
cis-1M-2E-Cyclopentane	398.31	401.20	0.7787	0.7896	1.4233	1.4269	a
trans-1M-2E-Cyclopentane	398.64	394.35	0.7791	0.7734	1.4220	1.4195	a
cis-1M-3E-Cyclopentane	389.67	394.26	0.7749	0.7712	1.4216	1.4170	a
trans-1M-3E-Cyclopentane	400.03	394.26	0.7784	0.7712	1.4215	1.4170	a
1,1,2MMM-Cyclopentane	385.86	386.88	0.7800	0.7771	1.4223	1.4205	a
1,1,3MMM-Cyclopentane	387.65	378.04	0.7732	0.7528	1.4204	1.4087	a
1,c-2,c-3MMM-Cyclopentane	386.39	396.15	0.7660	0.7837	1.4233	1.4238	a
1,c-2,t-3MMM-Cyclopentane	391.92	390.65	0.7816	0.7750	1.4237	1.4194	a
1,t-2,c-3MMM-Cyclopentane	388.16	383.35	0.7747	0.7581	1.4201	1.4114	a
1,c-2,c-4MMM-Cyclopentane	391.55	390.15	0.7742	0.7760	1.4217	1.4200	a

1,c-2,t-4MMM-Cyclopentane	394.37	389.88	0.7784	0.7680	1.4228	1.4161	a
1,t-2,c-4MMM-Cyclopentane	394.19	382.44	0.7783	0.7518	1.4218	1.4081	a
nB-Cyclopentane	435.09	429.75	0.7840	0.7893	1.4271	1.4293	a
IsoB-Cyclopentane	424.14	421.10	0.7908	0.7853	1.4286	1.4273	a
1M-1Propyl-Cyclopentane	427.47	419.15	0.7969	0.8036	1.4308	1.4350	a
1,1EE-Cyclopentane	424.59	423.65	0.8156	0.8072	1.4402	1.4363	a
cis-1,2EE-Cyclopentane	423.59	426.71	0.7870	0.8004	1.4279	1.4330	a
1,1MM-2E-Cyclopentane	416.47	411.15	0.7858	0.7928	1.4260	1.4300	a
nC5-CycPentane	459.11	453.65	0.7919	0.7954	1.4320	1.4336	a
nC6-LCycPentane	480.90	476.05	0.7995	0.8006	1.4371	1.4370	a
nC7-CycPentane	500.94	497.05	0.8042	0.8051	1.4399	1.4400	a
nC8-CycPentane	519.74	516.65	0.8091	0.8088	1.4436	1.4425	a
nC9-CycPentane	538.85	535.15	0.8129	0.8121	1.4468	1.4446	a
nC10-CycPentane	553.17	552.53	0.8162	0.8149	1.4478	1.4466	a
nC11CycPentane	570.19	568.95	0.8187	0.8175	1.4498	1.4482	a
nC12-CycPentane	584.88	584.35	0.8204	0.8197	1.4520	1.4497	a
nC13-CycPentane	599.75	599.05	0.8244	0.8217	1.4537	1.4510	a
nC14-CycPentane	612.60	613.15	0.8250	0.8235	1.4545	1.4522	a
nC15-CycPentane	625.38	626.15	0.8278	0.8252	1.4552	1.4533	a
nC16-CycPentane	637.46	639.15	0.8276	0.8267	1.4561	1.4543	a
nC17-CycPentane	649.81	650.15	0.8303	0.8280	1.4580	1.4552	a
nC18-CycPentane	662.04	662.15	0.8317	0.8293	1.4585	1.4560	a
nC19-CycPentane	671.22	673.15	0.8324	0.8303	1.4584	1.4568	a
nC20-CycPentane	682.60	683.15	0.8327	0.8315	1.4591	1.4575	a
CycloHexane	351.33	353.87	0.7705	0.7823	1.4226	1.4235	a
M-CycloHexane	378.56	374.08	0.7720	0.7748	1.4194	1.4206	a
E-CycloHexane	409.74	404.95	0.7861	0.7926	1.4249	1.4307	a
1,1MM-CycloHexane	395.74	392.70	0.7849	0.7854	1.4276	1.4266	a
cis-1,2MM-CycloHexane	396.17	402.94	0.7912	0.8006	1.4314	1.4336	a
trans-1,2MM-CycloHexane	397.07	396.58	0.7877	0.7803	1.4269	1.4247	a
cis-1,3MM-CycloHexane	397.78	393.24	0.7804	0.7704	1.4239	1.4206	a
trans-1,3MM-CycloHexane	397.00	397.61	0.7787	0.7892	1.4224	1.4284	a
cis-1,4MM-CycloHexane	396.54	397.47	0.7752	0.7873	1.4224	1.4273	a
trans-1,4MM-CycloHexane	400.33	392.51	0.7824	0.7670	1.4262	1.4185	a
Propyl-CycloHexane	436.12	429.90	0.7917	0.7981	1.4293	1.4348	a
IsoP-CycloHexane	423.75	427.91	0.8016	0.8064	1.4356	1.4386	a
nB-CycloHexane	459.54	454.13	0.7939	0.8033	1.4336	1.4385	a
IsoB-CycHexane	446.09	444.44	0.7945	0.8161	1.4318	1.4364	e
secB-CycloHexane	450.70	452.43	0.8066	0.8172	1.4386	1.4445	a
terB-CycloHexane	438.91	444.72	0.8093	0.8167	1.4412	1.4447	a

1M-4IsoP-CycloHexane	442.77	443.87	0.8047	0.8186	1.4374	1.4413	e
nC5-CycloHexane	482.22	476.85	0.8190	0.8077	1.4355	1.4416	a
nC6-CycloHexane	501.24	497.85	0.8219	0.8115	1.4382	1.4441	a
nC7-CycloHexane	520.27	518.05	0.8122	0.8148	1.4444	1.4463	a
nC8-CycloHexane	537.19	536.95	0.8142	0.8177	1.4464	1.4484	a
nC9-CycloHexane	553.50	554.65	0.8179	0.8202	1.4486	1.4499	a
nC10-CycloHexane	570.72	570.75	0.8323	0.8223	1.4552	1.4514	a
nC11-CycloHexane	583.26	586.25	0.8227	0.8244	1.4512	1.4527	a
nC12-CycloHexane	596.18	600.85	0.8223	0.8261	1.4516	1.4539	a
nC13-CycloHexane	609.26	614.65	0.8242	0.8277	1.4524	1.4550	a
nC14-CycloHexane	624.69	627.15	0.8358	0.8291	1.4583	1.4559	a
nC15-CycloHexane	638.09	640.15	0.8302	0.8303	1.4574	1.4568	a
nC16-CycloHexane	650.14	652.15	0.8318	0.8316	1.4582	1.4576	a
nC17-CycloHexane	660.37	664.15	0.8324	0.8327	1.4579	1.4583	a
nC18-CycloHexane	672.98	675.15	0.8407	0.8337	1.4626	1.4590	a
nC19-CycloHexane	682.86	685.15	0.8339	0.8346	1.4596	1.4596	a
nC20-CycloHexane	694.49	695.15	0.8408	0.8355	1.4622	1.4602	a
cis-Decalin	455.16	468.97	0.8816	0.9018	1.4716	1.4788	a
trans-Decalin	458.75	460.46	0.8861	0.8755	1.4738	1.4671	a
BiCycloHexyl	499.29	512.19	0.8911	0.8900	1.4792	1.4777	a
9E-cis-Decalin	507.80	505.93	0.8805	0.8900	1.4701	1.4780	a
9E-trans-Decalin	507.66	498.15	0.8741	0.8648	1.4672	1.4640	a
9M-trans-decalin	479.53	478.15	0.8807	0.8658	1.4695	1.4619	d
1,10MM-cis-decalin	494.84	493.15	0.9014	0.8936	1.4794	1.4790	d
1,10MM-trans-decalin	496.88	486.15	0.8714	0.8672	1.4657	1.4637	d
Saturate Test Set (N=34)							
E-CycloButane	340.01	343.75	0.7504	0.7327	1.4023	1.4000	a
1,1,2MMM-Cycloheptane	442.97	442.95	0.8058	0.8280	1.4391	1.4508	b
nC21-CycloPentane	693.54	693.15	0.8336	0.8323	1.4601	1.4583	b
nC22-CycloPentane	704.78	706.15	0.8351	0.8332	1.4606	1.4589	b
nC26-CycloPentane	742.63	741.15	0.8363	0.8363	1.4612	1.4609	b
B(3M)-Cyclohexane	470.35	469.65	0.8128	0.8059	1.4414	1.4401	b
nC23-CycloHexane	724.20	732.15	0.8376	0.8378	1.4611	1.4618	b
nC24-CycloHexane	726.67	740.15	0.8493	0.8384	1.4679	1.4622	b
nC25-CycloHexane	742.56	749.15	0.8507	0.8390	1.4684	1.4626	b
nC26-CycloHexane	751.18	757.15	0.8368	0.8396	1.4613	1.4630	b
nC28-CycloHexane	767.65	772.15	0.8530	0.8407	1.4692	1.4637	b
nC30-CycloHexane	786.19	787.15	0.8390	0.8416	1.4627	1.4643	b
nC34-CycloHexane	825.46	813.15	0.8594	0.8432	1.4626	1.4654	b
DicycloPentane	403.81	405.15	0.8818	0.8642	1.4670	1.4606	b
1cyclohexyl-3cyclopentyl-Propane	533.66	543.15	0.8844	0.8642	1.4785	1.4606	b
1,2-Dicyclohexyl-Ethane	534.31	545.65	0.8871	0.8781	1.4795	1.4743	b
1,2-Dicyclohexyl-Propane	559.57	557.65	0.8753	0.8763	1.4730	1.4771	b
2,2-Dicyclohexyl-Propane	551.06	559.15	0.9056	0.9053	1.4813	1.4897	b

1,1'-Dicyclohexyl-Butane	581.39	568.15	0.8699	0.8842	1.4684	1.4791	b
1,4-Dicyclohexyl-Butane	567.85	578.15	0.8887	0.8810	1.4820	1.4731	b
np-C31	731.92	732.38	0.8111	0.8147	1.4508	1.4523	d
np-C32	740.62	740.68	0.8121	0.8160	1.4511	1.4530	d
np-C33	749.81	749.20	0.8131	0.8172	1.4515	1.4536	d
np-C34	755.31	756.44	0.8112	0.8184	1.4504	1.4542	d
np-C35	767.06	763.93	0.8143	0.8195	1.4523	1.4548	d
np-C36	772.46	771.18	0.8117	0.8205	1.4506	1.4554	d
np-C37	783.66	778.20	0.8155	0.8215	1.4527	1.4559	d
np-C38	790.88	784.99	0.8155	0.8224	1.4521	1.4564	d
np-C39	794.99	791.58	0.8131	0.8233	1.4513	1.4568	d
np-C40	803.50	798.20	0.8134	0.8242	1.4511	1.4573	d
Dodecahydro-Fluorene	528.91	529.15	0.9425	0.9489	1.4967	1.5012	b,e
Perhydro-Anthracene	546.80	543.15	0.9378	0.9581	1.4966	1.4972	a,e
Perhydro-Phenanthrene	545.61	545.15	0.9408	0.9447	1.4984	1.5003	b,e
Perhydro-chrysene	628.52	626.15	0.9991	0.9858	1.5255	1.5215	b,e
Aromatic Training Set (N=200)							
BENZENE	351.90	353.24	0.8656	0.8832	1.4906	1.4979	a
TOLUENE	377.10	383.78	0.8643	0.8741	1.4970	1.4940	a
E-BENZENE	407.04	409.35	0.8644	0.8737	1.4899	1.4932	a
o-XYLENE	420.23	417.58	0.8843	0.8849	1.5055	1.5029	a
m-XYLENE	414.57	412.27	0.8726	0.8691	1.5004	1.4946	a
p-XYLENE	407.81	411.51	0.8628	0.8654	1.4959	1.4933	a
Propyl-LBENZENE	433.53	432.39	0.8620	0.8683	1.4900	1.4895	a
IsoP-BENZENE	424.68	425.56	0.8671	0.8682	1.4890	1.4889	a
o-ETHYLTOLUENE	436.30	438.33	0.8783	0.8851	1.4973	1.5021	a
m-ETHYLTOLUENE	435.02	434.48	0.8686	0.8692	1.4974	1.4941	a
p-ETHYLTOLUENE	433.85	435.16	0.8580	0.8655	1.4903	1.4924	a
1,2,3MMM-BENZENE	453.69	449.27	0.8968	0.8985	1.5090	1.5115	a
1,2,4MMM-BENZENE	444.00	442.53	0.8852	0.8805	1.5052	1.5024	a
1,3,5MMM-BENZENE	432.42	437.89	0.8795	0.8698	1.4976	1.4968	a
nB-BENZENE	457.50	456.46	0.8624	0.8660	1.4855	1.4874	a
IsoB-BENZENE	441.96	445.94	0.8599	0.8577	1.4823	1.4840	a
secB-BENZENE	459.71	446.48	0.8719	0.8657	1.4905	1.4878	a
terB-BENZENE	442.22	442.30	0.8757	0.8713	1.4902	1.4902	a
1M-2nP-BENZENE	459.31	457.95	0.8757	0.8780	1.4969	1.4974	a
1M-3nP-BENZENE	452.73	454.95	0.8637	0.8659	1.4948	1.4912	a
1M-4nP-BENZENE	454.47	456.45	0.8540	0.8637	1.4887	1.4898	a
o-CYMENE	449.21	451.33	0.8808	0.8812	1.4985	1.4983	a
m-CYMENE	450.13	448.23	0.8663	0.8655	1.4943	1.4905	a
p-CYMENE	450.84	450.28	0.8589	0.8608	1.4888	1.4885	a
o-EE-BENZENE	465.67	456.61	0.8792	0.8839	1.5000	1.5011	a
m-EE-BENZENE	462.87	454.29	0.8742	0.8683	1.4974	1.4931	a
p-EE-BENZENE	461.97	456.94	0.8582	0.8663	1.4914	1.4924	a
1,2MM-3E-BENZENE	466.51	467.11	0.8913	0.8966	1.5050	1.5095	a

1,2MM-4E-BENZENE	465.02	462.93	0.8777	0.8788	1.5000	1.5009	a
1,3MM-2E-BENZENE	459.40	463.19	0.8918	0.8948	1.5038	1.5085	a
1,3MM-4E-BENZENE	458.74	461.59	0.8771	0.8807	1.5022	1.5015	a
1,3MM-5E-BENZENE	450.61	456.93	0.8698	0.8692	1.4958	1.4958	a
1,4MM-2E-BENZENE	461.14	459.98	0.8783	0.8816	1.5017	1.5020	a
1,2,3,4MMMM-BENZENE	485.93	478.19	0.9091	0.9084	1.5151	1.5181	a
1,2,3,5MMMM-BENZENE	468.17	471.15	0.9006	0.8948	1.5098	1.5107	a
1,2,4,5MMMM-BENZENE	472.80	469.99	0.8892	0.8918	1.5070	1.5093	a
n-PENTYL-BENZENE	478.32	478.61	0.8606	0.8624	1.4824	1.4856	a
2-phenyl-pentane	474.61	466.15	0.8733	0.8623	1.4885	1.4853	d
3-phenyl-pentane	478.46	464.15	0.8806	0.8638	1.4947	1.4854	d
1-phenyl-2M-butane	475.57	470.15	0.8692	0.8628	1.4842	1.4840	d
1-phenyl-3M-butane	484.39	472.05	0.8520	0.8598	1.5005	1.4820	d
2-phenyl-2M-butane	480.17	465.53	0.8874	0.8787	1.4948	1.4935	d
2-phenyl-3M-butane	471.54	461.15	0.8750	0.8739	1.4882	1.4840	d
1-phenyl-2,2MM-propane	449.15	459.15	0.8685	0.8618	1.4808	1.4860	d
1M-2nB-benzene	480.89	481.15	0.8753	0.8749	1.4926	1.4940	d
1M-3nB-benzene	474.34	478.15	0.8621	0.8628	1.4905	1.4890	d
1M-4nB-benzene	475.22	480.15	0.8550	0.8608	1.4814	1.4880	d
1M-2secB-Benzene	481.19	469.15	0.8845	0.8769	1.4995	1.4950	d
1M-3secB-Benzene	477.86	467.15	0.8721	0.8618	1.4948	1.4880	d
1M-4secB-Benzene	477.06	470.15	0.8672	0.8699	1.4908	1.4910	d
1M-2isoB-Benzene	463.69	469.15	0.8700	0.8688	1.4889	1.4912	d
1M-3isoB-Benzene	465.93	467.15	0.8632	0.8574	1.4872	1.4865	d
1M-4isoB-Benzene	465.35	469.15	0.8576	0.8555	1.4838	1.4851	d
1M-2terB-Benzene	458.71	473.60	0.8881	0.8937	1.4978	1.5053	d
1M-3terB-Benzene	462.54	462.41	0.8747	0.8696	1.4935	1.4921	d
1M-4terB-Benzene	461.45	465.91	0.8693	0.8650	1.4881	1.4895	d
1E-2PropylBenzene	476.19	476.15	0.8786	0.8783	1.4969	1.4969	d
1E-3PropylBenzene	479.63	474.15	0.8652	0.8645	1.4970	1.4907	d
1E-4PropylBenzene	482.22	478.15	0.8604	0.8632	1.4918	1.4898	d
1E-2IsoP-Benzene	468.01	466.15	0.8820	0.8920	1.5005	1.5060	d
1E-3IsoP-Benzene	475.56	465.15	0.8714	0.8628	1.4989	1.4900	d
1E-4IsoP-Benzene	474.98	469.75	0.8639	0.8623	1.4916	1.4900	d
1,2MM-3PropylBenzene	488.40	483.85	0.8841	0.8904	1.5026	1.5053	d
1,2MM-4PropylBenzene	480.77	482.05	0.8721	0.8754	1.4991	1.4978	d
1,3MM-2PropylBenzene	475.56	480.75	0.8841	0.8895	1.5028	1.5041	d
1,3MM-4PropylBenzene	479.74	479.75	0.8702	0.8762	1.4971	1.4976	d
1,3MM-5PropylBenzene	468.92	475.39	0.8676	0.8645	1.4938	1.4930	d
1,4MM-2PropylBenzene	474.62	477.45	0.8756	0.8756	1.4983	1.4977	d
1,2MM-3IsoP-Benzene	479.76	475.75	0.8912	0.8920	1.5041	1.5060	d

1,2MM-4IsoP-Benzene	479.05	474.95	0.8758	0.8738	1.4983	1.4971	d
1,3MM-2IsoP-Benzene	469.44	472.15	0.8875	0.8940	1.5025	1.5070	d
1,3MM-4IsoP-Benzene	471.40	472.25	0.8773	0.8769	1.4996	1.4980	d
1,3MM-5IsoP-Benzene	461.64	467.65	0.8685	0.8658	1.4933	1.4930	d
1,4MM-2IsoP-Benzene	470.91	469.35	0.8799	0.8777	1.5021	1.4988	d
1M-2,3EE-Benzene	476.11	479.75	0.8916	0.8950	1.5066	1.5083	d
1M-2,4EE-Benzene	482.54	478.15	0.8782	0.8787	1.5037	1.5005	d
1M-2,5EE-Benzene	484.19	480.25	0.8749	0.8797	1.5016	1.5012	d
1M-2,6EE-Benzene	488.81	481.95	0.8877	0.8947	1.5072	1.5084	d
1M-3,4EE-Benzene	478.21	476.75	0.8788	0.8801	1.5028	1.5017	d
1M-3,5EE-Benzene	477.63	473.85	0.8704	0.8668	1.5007	1.4947	d
1,2,3MMM-4E-Benzene	498.85	493.55	0.8976	0.9059	1.5122	1.5158	d
1,2,3MMM-5E-Benzene	487.62	488.95	0.8906	0.8903	1.5072	1.5079	d
1,2,4MMM-3E-Benzene	491.67	489.75	0.8988	0.8990	1.5094	1.5111	d
1,2,4MMM-5E-Benzene	482.08	486.15	0.8850	0.8869	1.5038	1.5053	d
1,2,4MMM-6E-Benzene	484.79	486.15	0.8918	0.8937	1.5068	1.5096	d
1,3,5MMM-2E-Benzene	476.28	485.55	0.8865	0.8869	1.5053	1.5052	d
PentaMethyl-Benzene	506.49	504.95	0.9121	0.9211	1.5164	1.5250	d
n-HEXYLBENZENE	497.44	499.26	0.8602	0.8622	1.4801	1.4842	a
1,3-DiIsoP-BENZENE	492.89	476.33	0.8694	0.8629	1.4975	1.4875	a
1,4-DiIsoP-BENZENE	494.69	483.65	0.8639	0.8606	1.4902	1.4876	a
n-HEPTYLBENZENE	515.58	519.25	0.8580	0.8617	1.4783	1.4832	a
n-OCTYLBENZENE	533.07	537.55	0.8589	0.8602	1.4773	1.4824	a
n-NONYLBENZENE	548.94	555.20	0.8564	0.8596	1.4761	1.4817	a
n-DECYLBENZENE	564.90	571.04	0.8565	0.8590	1.4755	1.4811	a
n-UNDECYLBENZENE	580.06	586.40	0.8550	0.8587	1.4752	1.4807	a
n-DODECYLBENZENE	594.75	600.76	0.8559	0.8595	1.4749	1.4803	a
n-TRIDECYLBENZENE	608.48	614.43	0.8547	0.8584	1.4748	1.4800	a
n-TetradecylBenzene	625.73	627.15	0.8545	0.8587	1.4750	1.4797	a
n-PentadecylBenzene	638.99	639.15	0.8522	0.8587	1.4750	1.4794	a
n-HexadecylBenzene	648.24	651.15	0.8549	0.8586	1.4758	1.4792	a
n-heptadecylbenzene	658.32	658.85	0.8562	0.8580	1.4758	1.4790	e
1-phenyl-nonadecane	684.14	680.35	0.8551	0.8578	1.4765	1.4786	e
1-phenyl-tetracosane	734.30	727.45	0.8596	0.8569	1.4809	1.4780	e
NAPHTHALENE	496.86	491.14	1.0649	1.0281°	1.6244	1.6232	a
1M-NAPHTHALENE	513.05	517.83	1.0374	1.0242	1.6158	1.6151	a
2M-NAPHTHALENE	516.03	514.26	1.0292	1.0082	1.6152	1.6019	a
1E-NAPHTHALENE	536.07	531.48	1.0137	1.0115	1.6067	1.6040	a
2E-NAPHTHALENE	536.95	531.05	1.0123	0.9961	1.6032	1.5977	a
1,2MM-NAPHTHALENE	543.43	539.45	1.0219	1.0219	1.6109	1.6143	a
1,4MM-NAPHTHALENE	523.61	540.45	1.0100	1.0208	1.6031	1.6114	a
1-n-PropylNaphthalene	542.71	545.93	0.9937	0.9943	1.5938	1.5930	a
2- n-PropylNaphthalene	549.90	546.65	0.9927	0.9808	1.5927	1.5850	a
1-n-ButylNaphthalene	574.93	562.54	0.9783	0.9805	1.5760	1.5797	a
2-n-ButylNaphthalene	568.52	561.15	0.9759	0.9698	1.5799	1.5747	a
1-n-PentylNaphthalene	577.15	579.15	0.9738	0.9705	1.5743	1.5704	a
1-n-HexylNaphthalene	597.32	595.15	0.9616	0.9544	1.5676	1.5626	a

2-n-HexylNaphthalene	595.02	596.15	0.9575	0.9521	1.5631	1.5601	a
1-n-HeptylNaphthalene	607.36	610.15	0.9530	0.9537	1.5575	1.5565	a
2-n-HeptylNaphthalene	613.56	614.15	0.9509	0.9410	1.5592	1.5535	d
1-n-OctylNaphthalene	628.07	629.15	0.9445	0.9468	1.5544	1.5506	a
2-n-OctylNaphthalene	626.47	630.15	0.9405	0.9358	1.5508	1.5480	d
1-n-NonylNaphthalene	638.43	639.00	0.9377	0.9408	1.5477	1.5455	a
2-n-NonylNaphthalene	640.68	642.15	0.9337	0.9339	1.5462	1.5442	a
1-n-DecylNaphthalene	655.81	652.00	0.9328	0.9354	1.5445	1.5412	a
2-n-DrcylNaphthalene	658.05	660.15	0.9332	0.9253	1.5445	1.5392	d
1-nC11-Naphthalene	667.32	674.15	0.9286	0.9320	1.5407	1.5379	d
2-nC11-Naphthalene	668.31	674.15	0.9284	0.9254	1.5417	1.5356	d
1-nC12-Naphthalene	674.90	688.15	0.9251	0.9240	1.5356	1.5344	d
2-nC12-Naphthalene	679.76	687.15	0.9275	0.9177	1.5278	1.5323	d
BIPHENYL	532.83	528.15	1.0521	1.0323	1.5979	1.5873	a
1M-2-PHENYLBENZENE	540.14	528.45	1.0203	1.0159	1.5913	1.5890	a
1M-3-PHENYLBENZENE	542.67	545.85	1.0217	1.0185	1.5930	1.6016	a
DIPHENYLMETHANE	540.86	537.42	1.0142	1.0101	1.5740	1.5752	a
1,1-DiphenylETHANE	554.04	545.78	0.9987	1.0041	1.5692	1.5702	a
1,2-DiphenylETHANE	560.46	553.65	0.9903	0.9914	1.5633	1.5704	a
1,1-DiphenylPROPANE	560.64	556.37	0.9864	0.9910	1.5665	1.5620	a
1,2-DiphenylPROPANE	564.00	556.81	0.9812	0.9817	1.5568	1.5562	a
1,1-DiphenylBUTANE	571.59	567.44	0.9717	0.9793	1.5612	1.5546	a
1,1-DiphenylPENTANE	580.65	581.04	0.9650	0.9620	1.5538	1.5489	a
1,1-DiphenylHEXANE	596.22	594.18	0.9570	0.9605	1.5480	1.5428	a
1,1-DiphenylHEPTANE	606.80	607.15	0.9526	0.9542	1.5434	1.5381	a
1,1-DiphenylOCTANE	621.91	619.15	0.9444	0.9468	1.5388	1.5336	a
1,1-DiphenylNONANE	629.43	630.15	0.9399	0.9413	1.5338	1.5299	a
1,1-DiphenylDECANE	637.35	640.15	0.9344	0.9364	1.5282	1.5266	a
1,1-DiphenylUndecane	648.80	650.15	0.9288	0.9322	1.5245	1.5238	a
1,1-DiphenylDodecane	657.38	659.15	0.9213	0.9284	1.5190	1.5213	a
1,1-DiphenylTridecane	668.31	668.15	0.9213	0.9248	1.5158	1.5190	a
1,1-DiphenylTetradecane	676.73	676.15	0.9168	0.9224	1.5126	1.5182	a
1,1-DiphenylPentaDecane	684.11	684.15	0.9133	0.9190	1.5087	1.5151	a
1,1-DiphenylHexaDECANE	694.56	691.15	0.9084	0.9173	1.5048	1.5140	a
1,3MM-Naphthalene	533.39	536.15	1.0093	1.0108	1.6064	1.6068	d
1,6MM- Naphthalene	533.97	536.15	1.0129	1.0075	1.6076	1.6051	d
1,7MM- Naphthalene	537.15	536.15	1.0094	1.0075	1.6099	1.6050	d
1,3,7MMM-Naphthalene	551.09	553.15	0.9915	0.9845	1.5974	1.5989	c,e
2,3,5MMM-NAPHTHALENE	564.61	558.15	1.0054	1.0133	1.6063	1.6015	c,e
Anthracene	612.36	613.15	1.1493	1.2445°	1.7081	1.7290	a,e
1M-Anthracene	622.88	636.15	1.1173	1.1001	1.6964	1.7052	b,c
Phenanthrene	609.12	611.15	1.1661	1.1750	1.7062	1.6952	a,e
Indane	445.77	451.12	0.9858	0.9686	1.5351	1.5358	a

1-M-Indane	470.56	463.75	0.9713	0.9437	1.5331	1.5241	a
2-M-Indane	466.54	464.55	0.9680	0.9464	1.5242	1.5193	a
4-M-Indane	475.81	478.65	0.9791	0.9608	1.5367	1.5333	a
5-M-Indane	470.32	475.15	0.9657	0.9495	1.5338	1.5311	a
Tetralin	472.00	480.77	0.9719	0.9748	1.5330	1.5392	b,c
1-M-Tetralin	487.63	493.74	0.9605	0.9623	1.5273	1.5333	a
1-E-Tetralin	508.78	512.72	0.9577	0.9569	1.5288	1.5298	a
2,2-MM-Tetralin	500.78	503.15	0.9495	0.9404	1.5194	1.5180	a
2,6-MM-Tetralin	510.07	510.93	0.9434	0.9464	1.5251	1.5240	a
6,7-MM-Tetralin	514.01	525.15	0.9463	0.9584	1.5290	1.5360	a
1-nC3-Tetralin	522.60	529.55	0.9444	0.9480	1.5224	1.5255	a
6-nC3-Tetralin	527.59	536.15	0.9335	0.9401	1.5281	1.5241	a
1-nC4-Tetralin	542.62	546.28	0.9392	0.9382	1.5192	1.5198	a
6-nC4-Tetralin	546.17	554.15	0.9270	0.9334	1.5223	1.5210	a
1-nC5-Tetralin	558.29	562.78	0.9334	0.9310	1.5155	1.5158	a
6-nC5-Tetralin	562.26	570.15	0.9183	0.9277	1.5213	1.5168	a
1-nC6-Tetralin	575.05	578.15	0.9274	0.9251	1.5122	1.5127	a
1-nC7-Tetralin	590.07	594.15	0.9235	0.9203	1.5089	1.5101	a
1-nC8-Tetralin	605.10	608.15	0.9181	0.9161	1.5063	1.5080	a
1-nC9-Tetralin	619.02	621.15	0.9160	0.9124	1.5045	1.5061	a
1-nC10-Tetralin	633.00	634.15	0.9115	0.9093	1.5023	1.5045	a
2M-Tetralin	490.92	494.15	0.9594	0.9562	1.5250	1.5290	b
5M-Tetralin	494.07	507.15	0.9688	0.9763	1.5350	1.5419	b
6M-Tetralin	490.36	502.15	0.9539	0.9580	1.5314	1.5337	b
2E-Tetralin	507.68	508.15	0.9460	0.9401	1.5178	1.5210	b
5E-Tetralin	516.89	518.15	0.9584	0.9773	1.5354	1.5380	b
6E-Tetralin	516.60	517.15	0.9433	0.9644	1.5291	1.5374	b
1,1MM-Tetralin	497.67	494.15	0.9559	0.9542	1.5243	1.5272	b
1,2MM-Tetralin	515.75	508.15	0.9545	0.9512	1.5259	1.5266	b
1,3MM-Tetralin	498.53	507.15	0.9475	0.9442	1.5233	1.5230	b
1,4MM-Tetralin	502.41	499.15	0.9587	0.9442	1.5257	1.5230	b
1,5MM-Tetralin	511.39	512.15	0.9576	0.9452	1.5308	1.5240	b
2,3MM-Tetralin	523.67	505.15	0.9613	0.9442	1.5268	1.5210	b
2,5MM-Tetralin	512.69	509.15	0.9557	0.9502	1.5313	1.5240	b
2,7MM-Tetralin	512.40	510.15	0.9431	0.9452	1.5257	1.5240	b
2,8MM-Tetralin	511.27	509.15	0.9557	0.9452	1.5298	1.5240	b
5,6MM-Tetralin	522.70	525.15	0.9620	0.9793	1.5363	1.5500	b
5,7MM-Tetralin	509.74	526.15	0.9509	0.9626	1.5326	1.5385	b
5,8MM-Tetralin	510.90	527.15	0.9618	0.9713	1.5327	1.5450	b
1,1,6MMM-Tetralin	515.01	513.15	0.9409	0.9398	1.5228	1.5237	b
1,1,2,6MMMM-Tetralin	547.98	528.95	0.9445	0.9374	1.5234	1.5197	b
Aromatic Test Set (N=61)							
9-p-tolyl-Octadecane	606.66	604.25	0.8551	0.8587	1.4781	1.4792	b
n-Octadecyl Benzene	672.15	681.15	0.8549	0.8579	1.4762	1.4788	d
1-phenyl-nonacosane	780.46	766.65	0.8605	0.8562	1.4809	1.4775	d
1-phenyl-tetratriacontane	810.72	811.15	0.8610	0.8555	1.4862	1.4772	d
1-phenyl	822.70	817.15	0.8562	0.8543	1.4840	1.4772	d

pentatriacosane							
1-phenyl Hexatriaconsane	823.71	822.15	0.8573	0.8542	1.4853	1.4771	d
ACENAPHTHENE	550.92	552.15	1.1288	NA	1.6513	1.6420	b
FLUORENE	561.49	570.44	1.1331	1.1521	1.6367	1.6470	a,e
FLUORANTHENE	646.00	655.95	1.2375	1.2410	1.7273	1.7390	b
9,10dihydro-Phenanthrene	574.74	573.05	1.1001	1.0957	1.6297	1.6393	b
1-8Octahydro-Anthracene	563.18	567.15	1.0018	1.0031	1.5663	1.5572	b
1-8Octahydro-Phenanthrene	565.37	568.15	1.0167	1.0260	1.5708	1.5569	b
1,3,5EEE-Benzene	493.77	488.93	0.8734	0.8694	1.4999	1.4932	e
3-Octylundecyl-Benzene	658.26	669.45	0.8597	0.8598	1.4783	1.4787	b
1,2-DIPHENYL-[3,3MM]-Ethane	585.28	571.15	0.9631	0.9703	1.5519	1.5546	b
1-(phenylmethyl)-4nP-Benzene	582.59	566.55	0.9596	0.9763	1.5572	1.5534	b
1,1diBenzyl-Propane	583.48	577.65	0.9626	0.9777	1.5423	1.5510	b
1,1-Diphenyl-[4,4MM]-Dodecane	686.46	679.35	0.9121	0.9117	1.5104	1.4981	b
CyclohexylBenzene	507.68	513.27	0.9460	0.9475	1.5178	1.5239	a
dodecahydro-Phenanthrene	525.56	522.35	0.9806	0.9717	1.5017	1.5082	b
1-4'-9,10'octahydro-phenanthrene	559.73	561.75	1.0189	1.0017	1.5643	1.5507	b
1,4,8MMM-6IsoP-Naphthalene	583.42	583.85	0.9687	0.9801	1.5826	1.5840	e
1M-4-PhenylBenzene	550.68	543.15	1.0206	1.0109	1.5923	1.5934	e
1,2-DiPhenylBenzene	647.34	611.15°	1.0933	1.0820	1.6478	1.6405	a
Chrysene	710.81	704.15	1.2119	1.2013	1.7686	1.7850	a
Triphenylene	697.36	702.15	1.2182	1.1952	1.7665	1.7560	a,e
Pyrene	669.23	666.15	1.2539	1.2590	1.7922	1.7700	a,e
1-phenyl-nonatriacontane	831.03	828.45	0.8603	0.8550	1.4884	NA	a
1-phenyl-tetratetracontane	843.94	853.05	0.8469	0.8545	1.4875	NA	a
1,3-Diphenyl-IsoButane	574.84	576.15	0.9694	0.9669	1.5463	NA	b
1,3-Diphenyl-2E-Propane	587.70	577.65	0.9711	0.9734	1.5464	NA	b
1M-7IsoP-Phenanthrene	662.28	663.15	1.0500	1.0452	1.653	NA	e
1-4tetrahydro-Phenanthrene	583.23	587.25	1.0772	1.0801	1.635	NA	b
9E-9,10dihydro-Anthracene	593.19	594.65	1.0546	1.0480	1.6068	NA	b
1,3-Diphenyl-BENZENE	648.98	650.00	1.1099	1.0902	1.6589	NA	a
1,4-Diphenyl-BENZENE	660.46	657.15	1.1109	1.0996	1.6549	NA	a
9,10dihydro-Anthracene	576.13	578.15	1.0977	1.215°	1.6212	NA	b

perylene	769.35	770.15	1.2970 NA	1.8478 NA	c
Tetraphenylethylene	693.22	693.15	1.0152 NA	1.5723 NA	b
2M-Anthracene	630.81	632.15	1.1089 NA	1.6954 NA	c
2,6MM-Anthracene	644.79	643.15	1.0901 NA	1.6841 NA	c
2,7MM-Anthracene	636.14	643.15	1.0795 NA	1.6766 NA	c
1M-Phenanthrene	624.53	632.15	1.1192 NA	1.6941 NA	c
2M-Phenanthrene	625.12	628.15	1.1253 NA	1.693 NA	c
3M-Phenanthrene	625.79	625.15	1.1202 NA	1.6939 NA	c
3,6MM-Phenanthrene	637.38	636.15	1.1018 NA	1.6816 NA	c
2M-Pyrene	682.85	683.15	1.2106 NA	1.7761 NA	c
benzo(a)pyrene	773.74	769.15	1.2834 NA	1.8513 NA	f
benzo(e)pyrene	763.21	766.15	1.2794 NA	1.8441 NA	f
Naphthacene	715.48	723.15	1.2032 NA	1.7697 NA	f
picene	810.21	792.15	1.2662 NA	1.8236 NA	f
dibenz(a,h)anthracene	809.57	808.15	1.252 NA	1.8262 NA	f
dibenz(a,j)anthracene	809.09	804.15	1.2514 NA	1.8295 NA	f
dibenz(a,c)anthracene	795.01	808.15	1.2536 NA	1.827 NA	f
anthanthrene	831.80	820.15	1.3481 NA	1.9248 NA	f
benzo(ghi)perylene	827.80	815.15	1.3654 NA	1.9198 NA	f
dibenzo(a,h)pyrene	865.75	869.15	1.3174 NA	1.8989 NA	f
dibenzo(a,e)pyrene	860.84	865.15	1.312 NA	1.9015 NA	f
dibenzo(a,l)pyrene	855.99	868.15	1.3036 NA	1.8965 NA	f
dibenzo(a,i)pyrene	865.76	867.15	1.325 NA	1.9006 NA	f
coronene	864.46	863.15	1.4029 NA	1.9869 NA	f

Shorthand

M = Methyl
E = Ethyl
IsoP = Iso-propyl
nB = n-Butyl
IsoB = Iso-butyl
terB = ter-Bubyl
secB = sec-Butyl
nCi = Straight chain subsituent of i carbon atoms
nP-Ci = normal paraffin of i carbon atms
o-Outlier

Data Reference

a = API Technical data book
b = CRC handbook
c = Handbook of PAH
d = Selected Value of Physical and Thermodynamic Properties
e =Beilstein Database
f =Spectral Atlas of PAC

Appendix B. The derivation of uniqueness for quasi-equilibrium isomer distribution

Smith and Missen (1982) proved analytically that the chemical equilibrium problem has a unique solution for a single phase of ideal solution. Similarly, the solution uniqueness of the optimization problem in Equation 3-4 can be proved analytically under the same condition. The constrained optimization problem can be solved by the Lagrange Multiplier method. The Lagrangian function is defined as

$$F(x,\lambda) = \Delta G_n^0 + \lambda_1\left(1 - \sum x_i\right) + \lambda_2\left(BP_{lump} - \sum x_i BP_i\right) \qquad \text{(B-1)}$$

Minimizing F with respect to x and λ results in a set of nonlinear equations

$$\frac{\partial F}{\partial x_i} = \frac{\partial \Delta G_n^0}{\partial x_i} - \lambda_1 - \lambda_2 BP_i = 0 \qquad i = 1,2,\cdots,N \qquad \text{(B-2)}$$

and constraints: $\frac{\partial F}{\partial \lambda_j} = 0; \qquad j = 1,2 \qquad \text{(B-3)}$

Substituting the derivatives of Equation B-1 into Equation B-2 and solving for x_i, we obtain

$$x_i = e^{\frac{\lambda_1 + \lambda_2 BP_i - \Delta G_i^0 - RT}{RT}} \qquad \text{(B-4)}$$

Therefore, minimizing the Lagrangian function is reduced to solve the following set of nonlinear equations

$$\sum_{i=1}^{N} e^{\frac{\lambda_1 + \lambda_2 BP_i - \Delta G_i^0 - RT}{RT}} - 1 = 0 \qquad \text{(B-5)}$$

and

$$\sum_{i=1}^{N} BP_i \cdot e^{\frac{\lambda_1+\lambda_2 BP_i-\Delta G_i^0-RT}{RT}} - BP_{lump} = 0 \qquad \text{(B-6)}$$

Mathematically, Equations B-5 and B-6 can be written as

$$F_1 = \sum \exp(a_i + t_1 + b_i t_2) - 1 \qquad \text{(B-7)}$$

$$F_2 = \sum b_i \cdot \exp(a_i + t_1 + b_i t_2) - c \qquad \text{(B-8)}$$

where $t_1 = \lambda_1/RT$, $t_2 = \lambda_2/RT$, $a_i = -(1+\Delta G^0_i/RT)$, $b_i = BP_i$, $c = BP_{lump}$.

The solution to the optimization problem is $(\lambda_1, \lambda_2) = f(T, t_1, t_2)$, such that

$$\underline{F}(t) = \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = 0 \qquad \text{(B-9)}$$

Assuming Z_1 and Z_2 are two roots of $\underline{F}(t)$:

$$F(Z_1) - F(Z_2) = F'(\xi)(Z_1 - Z_2) = 0 \qquad \text{(B-10)}$$

Let:

$$Z_1 - Z_2 = \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \qquad \text{(B-11)}$$

If multiple solutions exist, Equations B-10 and B-11 give

$$\begin{pmatrix} \sum e^{a_i+\xi_1+b_i\xi_2} & \sum b_i e^{a_i+\xi_1+b_i\xi_2} \\ \sum b_i e^{a_i+\xi_1+b_i\xi_2} & \sum b_i^2 e^{a_i+\xi_1+b_i\xi_2} \end{pmatrix} \begin{pmatrix} m_1 \\ m_2 \end{pmatrix} = 0 \qquad \text{(B-12)}$$

Then

$$\sum m_1 e^{a_i+\xi_1+b_i\xi_2} + \sum m_2 b_i e^{a_i+\xi_1+b_i\xi_2} = 0 \qquad \text{(B-13)}$$

and

$$\sum m_1 b_i e^{a_i+\xi_1+b_i\xi_2} + \sum m_2 b_i^2 e^{a_i+\xi_1+b_i\xi_2} = 0 \qquad \text{(B-14)}$$

With the stoichiometric constraint $\sum_{i=1}^{N} e^{a_i+\xi_1+b_i\xi_2} = 1$, Equation B-13 can be rearranged to give

$$m_1 = -m_2 \sum b_i e^{a_i+\xi_1+b_i\xi_2} \qquad \text{(B-15)}$$

Substituting Equation B-15 into B-14 yields

$$m_2 \sum b_i^2 e^{a_i+\xi_1+b_i\xi_2} - m_2 (\sum b_i e^{a_i+\xi_1+b_i\xi_2})^2 = 0 \qquad \text{(B-16)}$$

Physically, $0 < exp(a_i+\xi_1+b_i\xi_2) < 1$; and $b_i = BP_i$ (K)>0 for all hydrocarbons. Thus,

$$\sum b_i^2 e^{a_i+\xi_1+b_i\xi_2} \neq (\sum b_i e^{a_i+\xi_1+b_i\xi_2})^2$$

Therefore, to satisfy Equations B-15 and B-16, the following equation has to be true

$$m_1 = m_2 = 0 \qquad \text{(B-17)}$$

One can conclude, then, that multiple solutions do not exist. The nonlinear Equations B-5 and B-6 have a unique solution of the Lagrange multiplier (λ_1 and λ_2). Consequently, the composition x_i determined by Equation B-4 is also unique.

Appendix C. Figures C1-C14. Comparisons of predicted distributions of haptane isomers with the reported data in 18 crude oils (Martin et al., 1963)

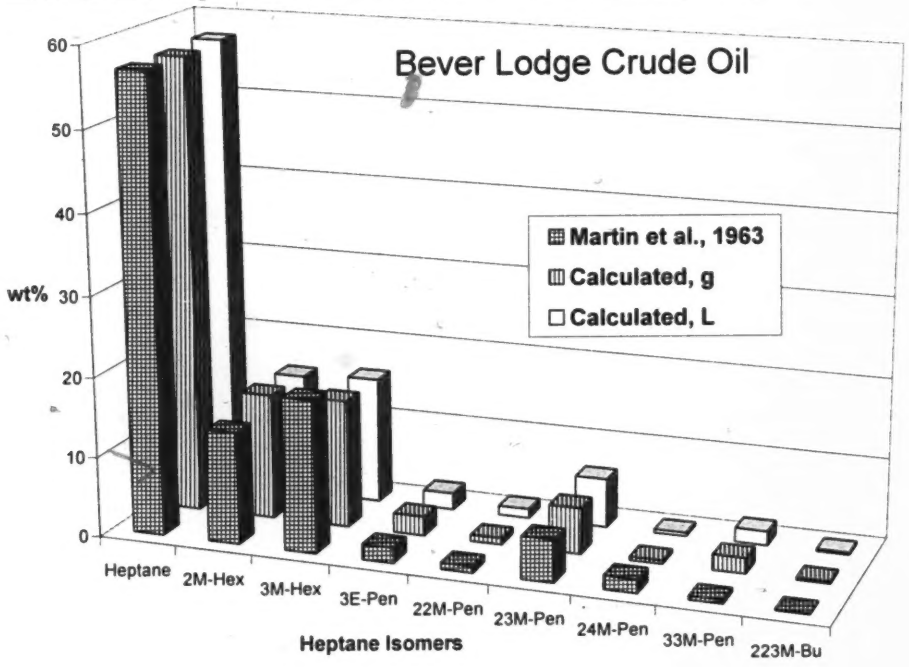
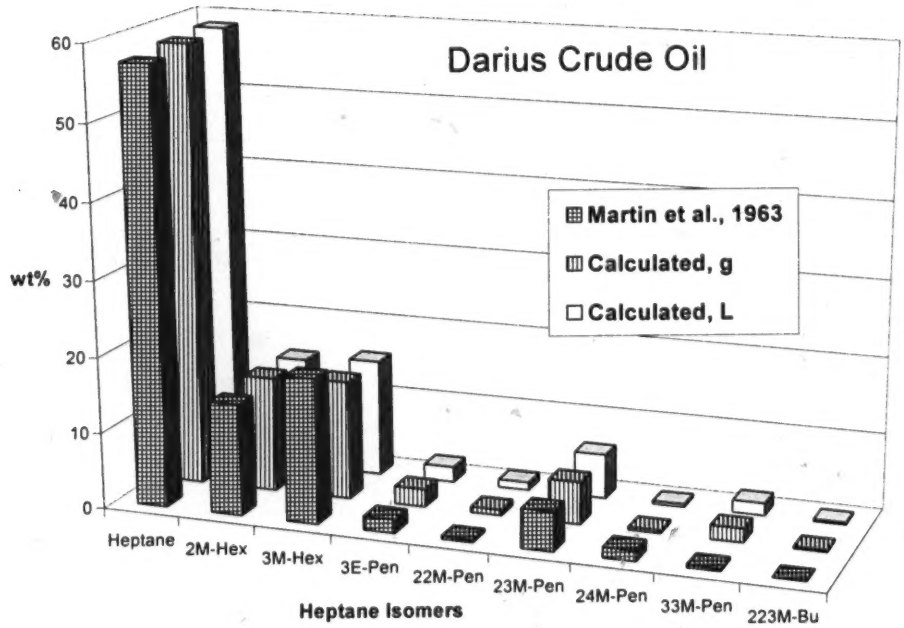


Figure C1. Prediction of heptane isomer distribution in Bever Lodge crude oil



FigureC2. Prediction of heptane isomer distribution in Darius crude oil

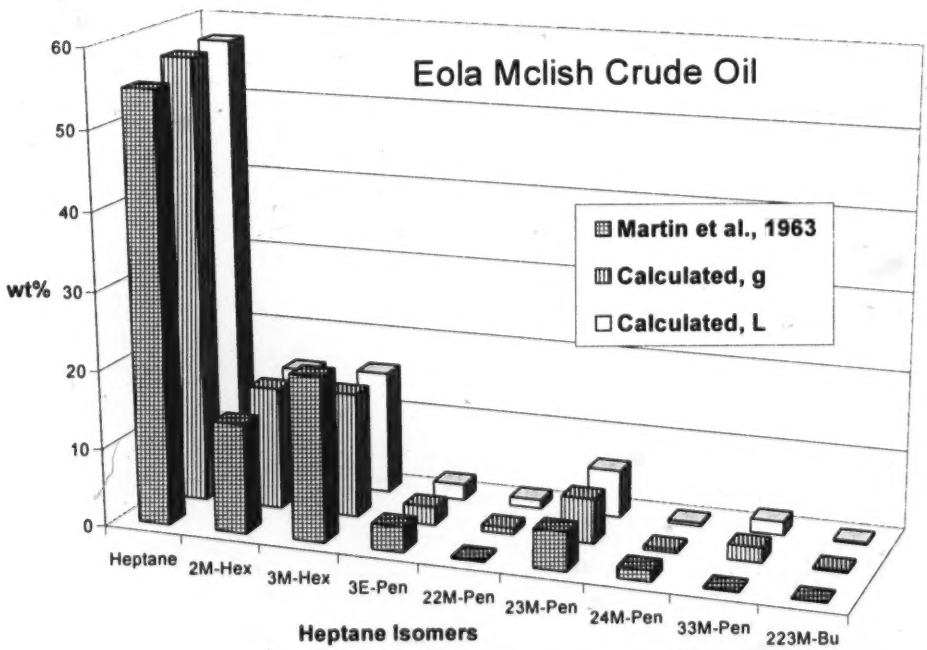


Figure C3. Prediction of heptane isomer distribution in Eola Mclish crude oil

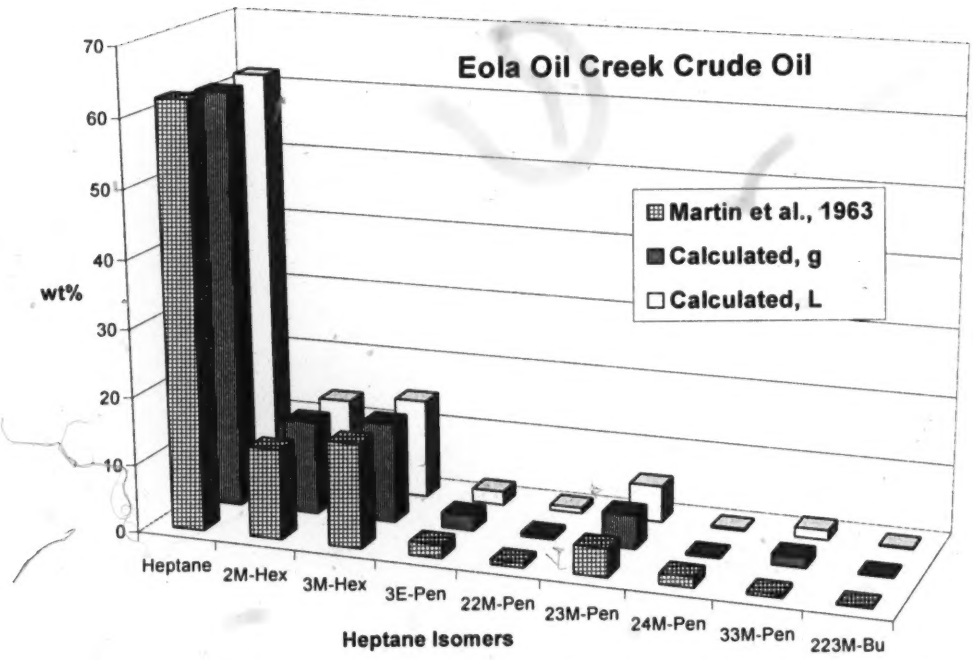


Figure C4. Prediction of heptane isomer distribution in Eola Oil Creek crude oil

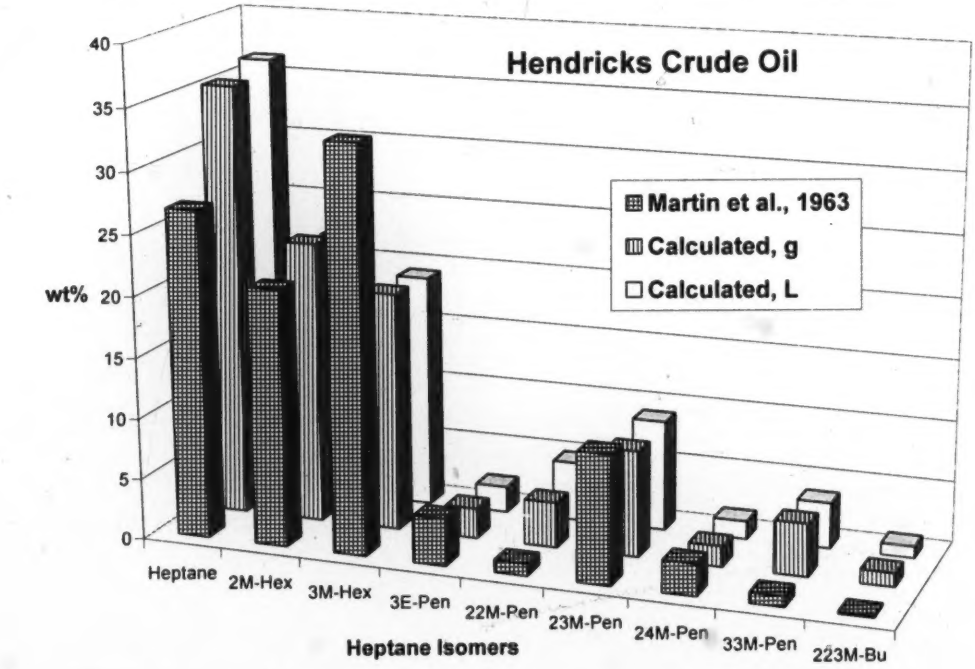


Figure C5. Prediction of heptane isomer distribution in Hendricks crude oil

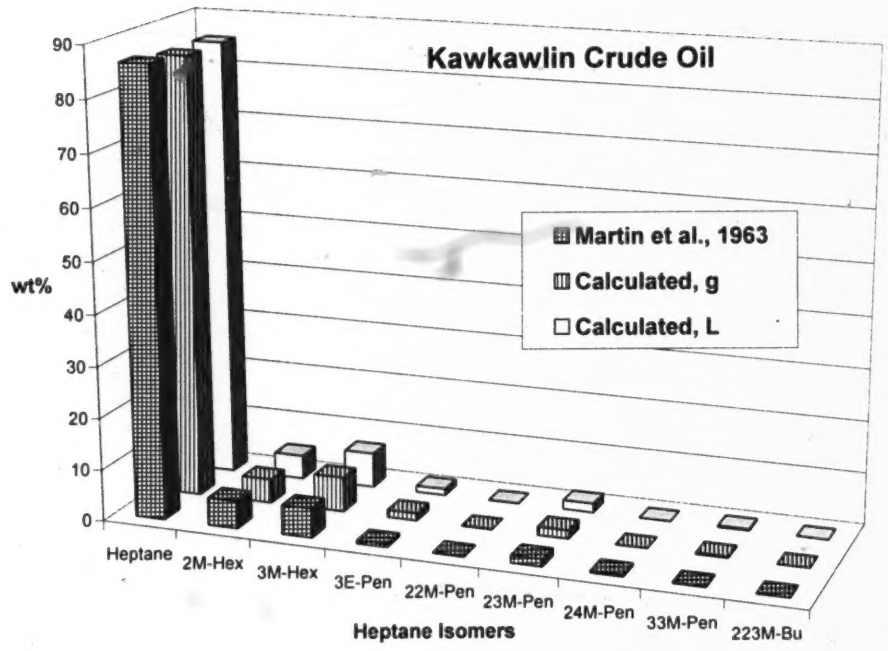


Figure C6. Prediction of heptane isomer distribution in Kawkawlin crude oil

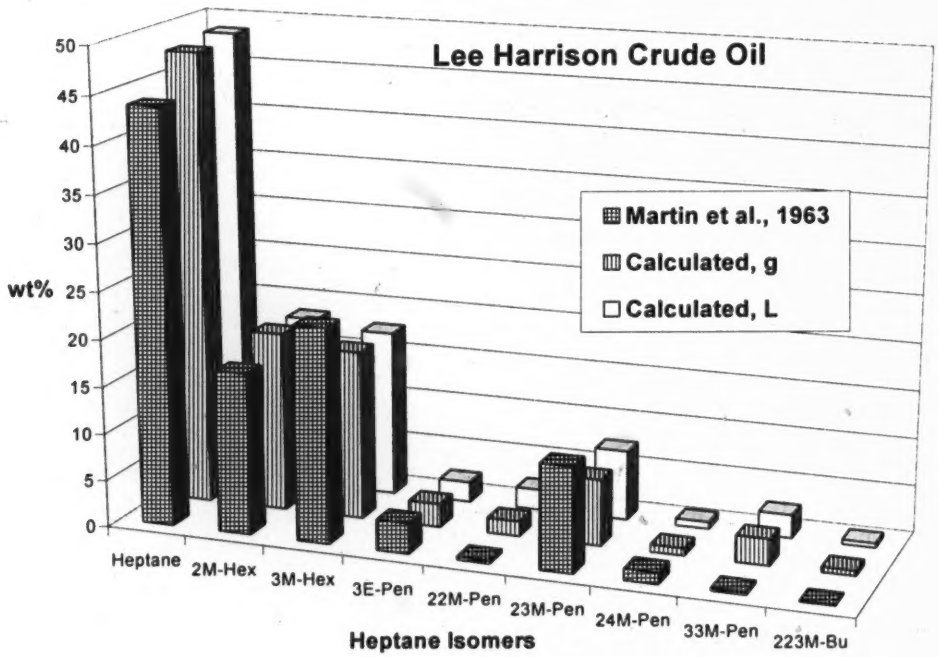


Figure C7. Prediction of heptane isomer distribution in Lee Harrison crude oil

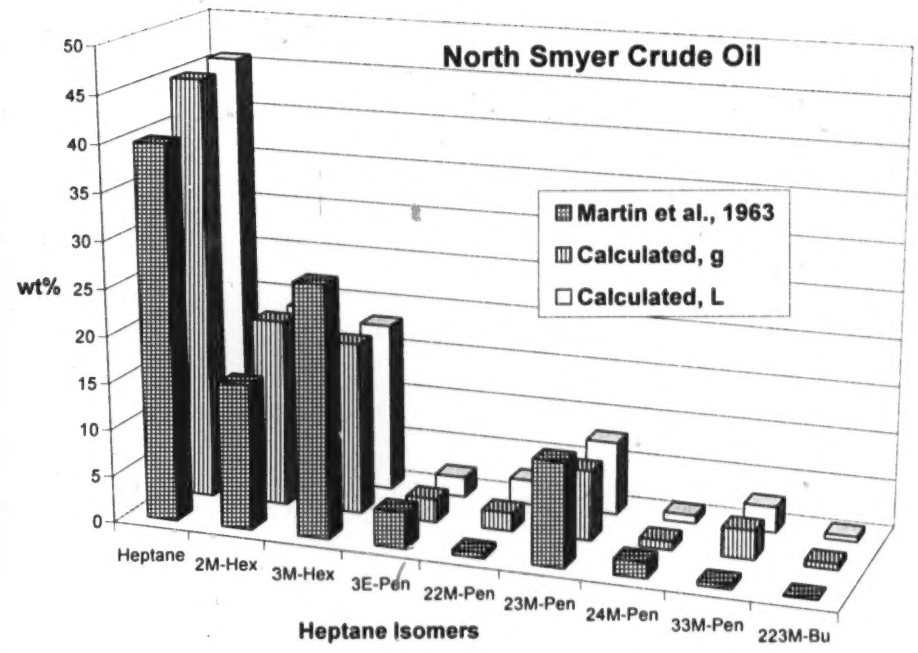


Figure C8. Prediction of heptane isomer distribution in North Smyer crude oil

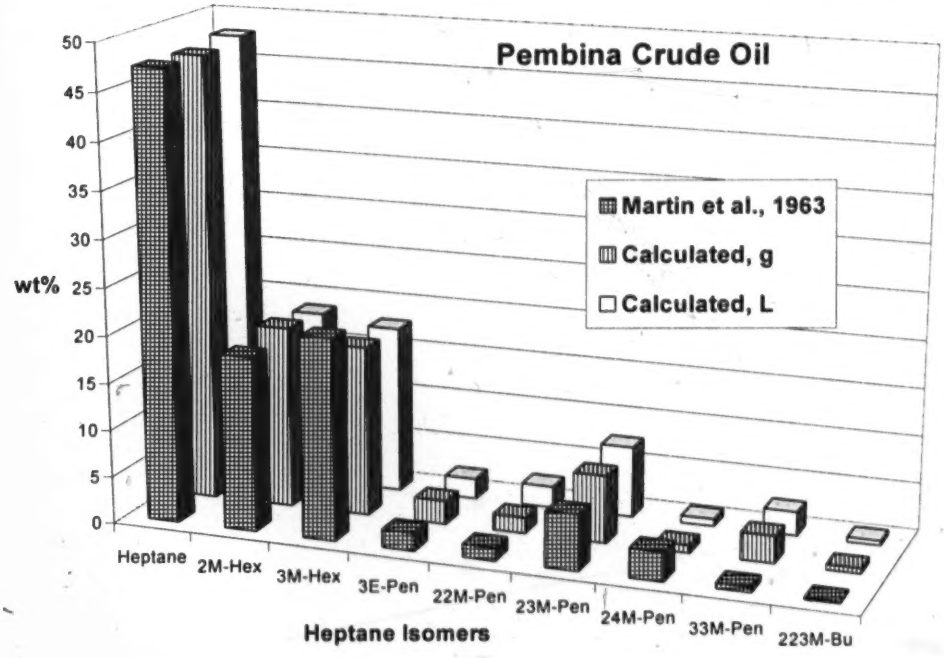


Figure C9. Prediction of heptane isomer distribution in Pembina crude oil

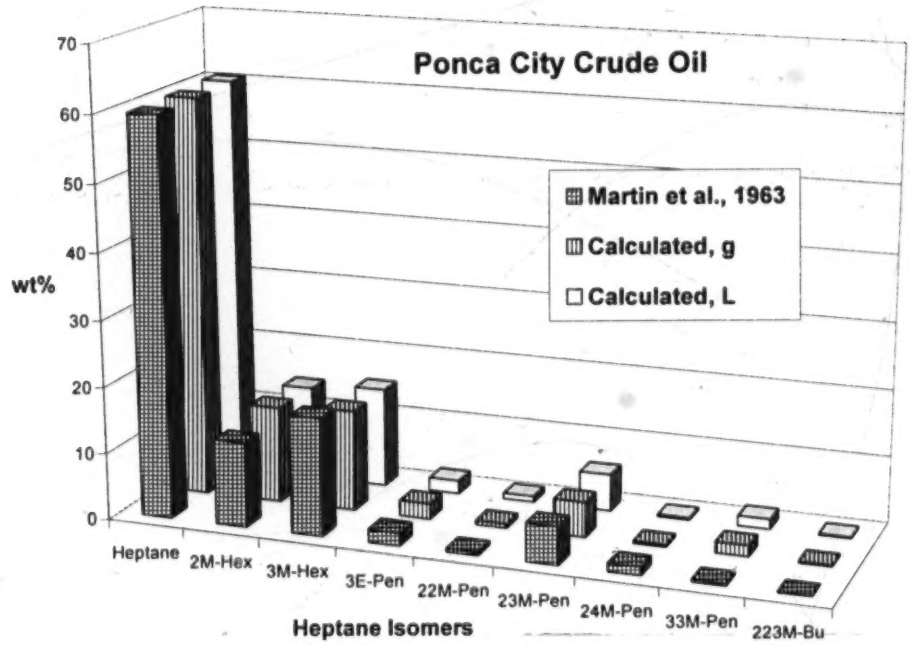


Figure C10. Prediction of heptane isomer distribution in Ponca City crude oil

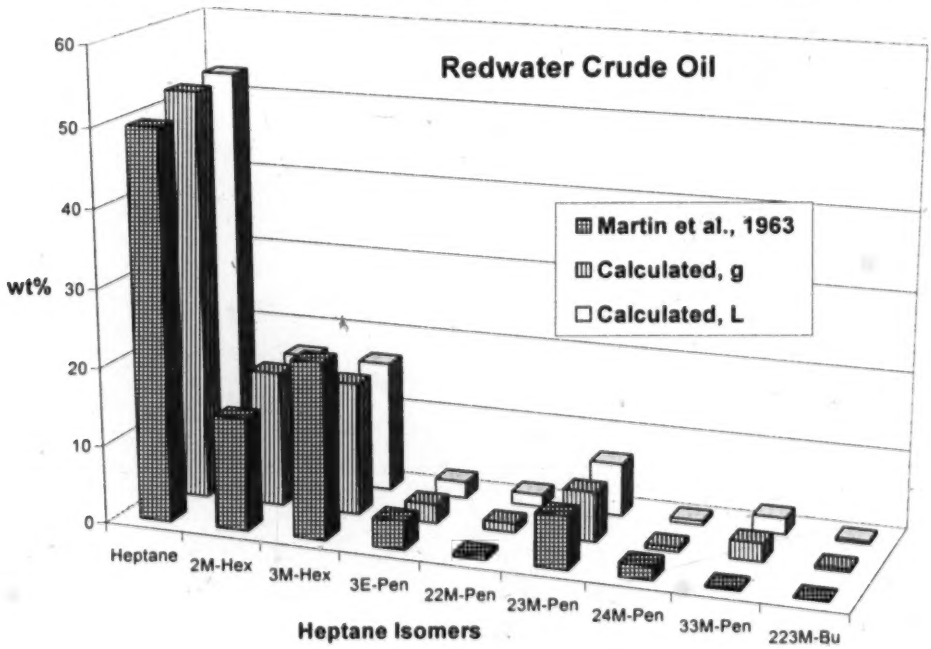


Figure C11. Prediction of heptane isomer distribution in Redwater crude oil

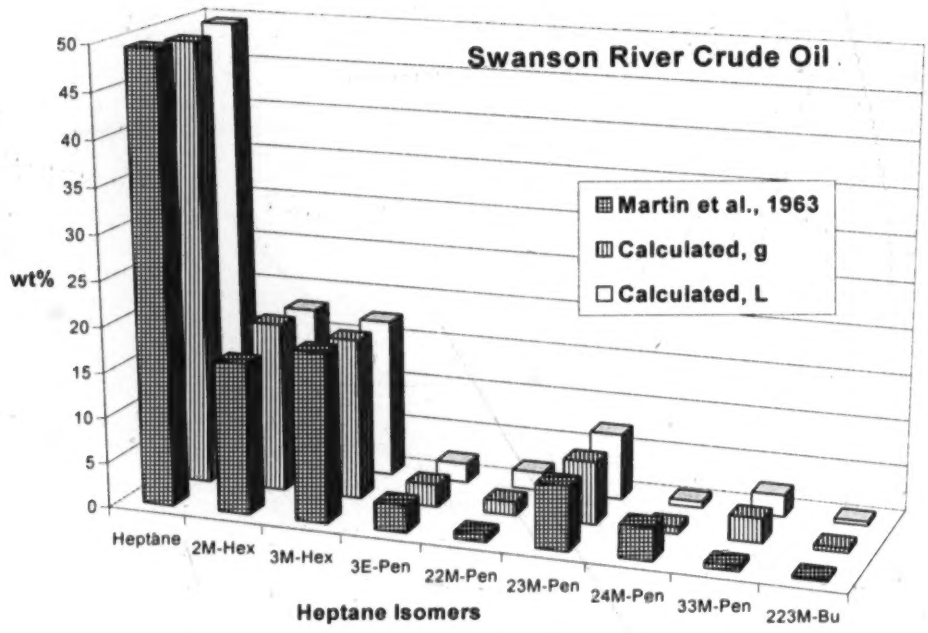


Figure C12. Prediction of heptane isomer distribution in Swanson River crude oil

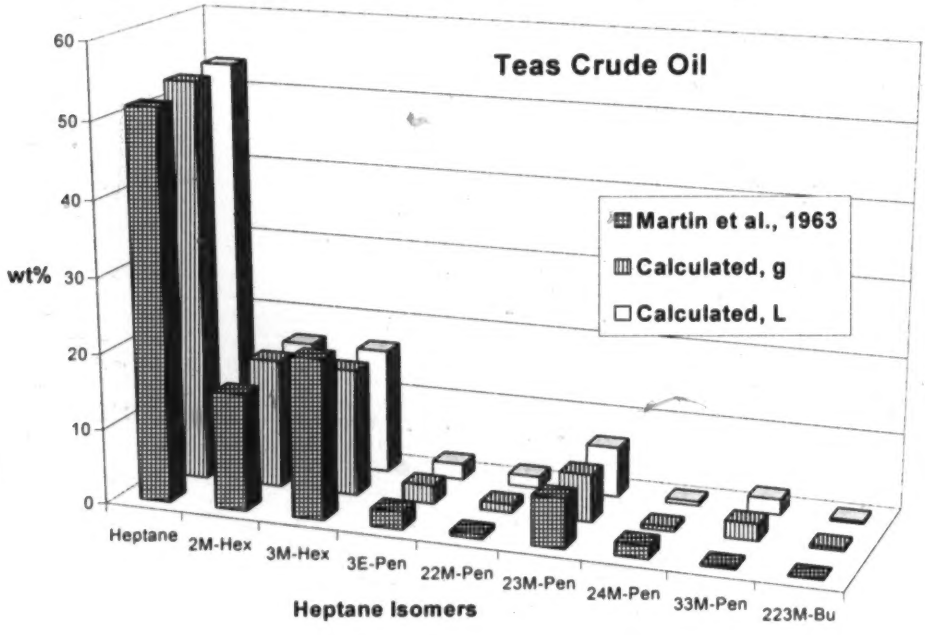


Figure C13. Prediction of heptane isomer distribution in Teas crude oil

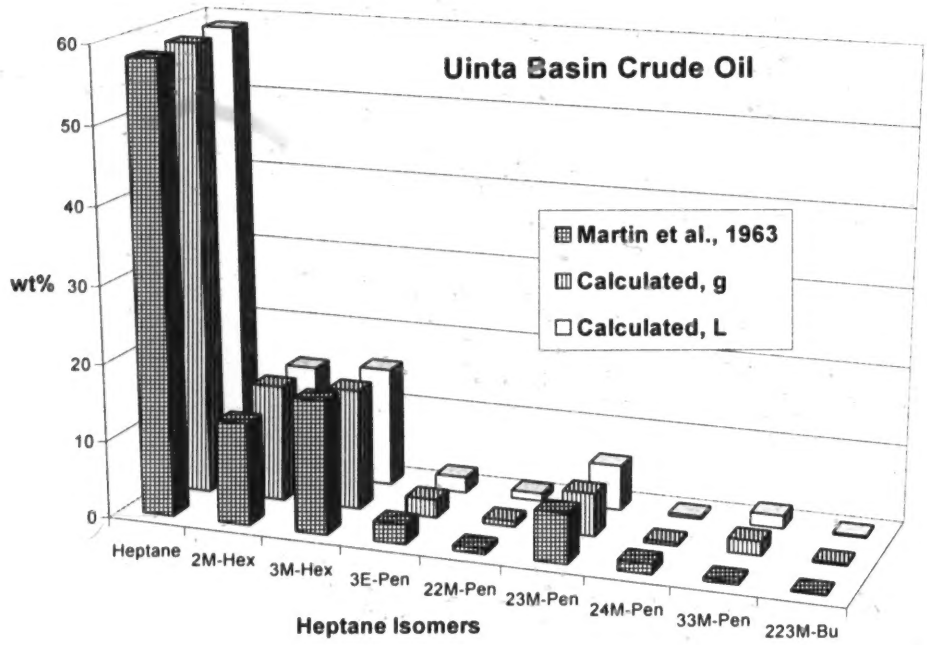


Figure C14. Prediction of heptane isomer distribution in Uinta Basin crude oil

Appendix D. PIONA, GC-FIMS, and data reconciliation results of sample 2-5

Table D1. PIONA report for sample 2 (<200°C fractions)

#C	Naphthenes	Iso-Paraffins	n-Paraffins	Aromatics	Totals
3	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.002	0.000	0.002
6	0.013	0.003	0.007	0.005	0.028
7	0.107	0.017	0.028	0.052	0.203
8	0.327	0.122	0.140	0.358	0.947
9	0.515	0.447	0.340	0.892	2.194
10	0.558	0.740	0.577	3.091	4.966
11	0.232	1.057	0.892	0.000	2.181
Totals	1.763	2.548	2.178	4.031	10.521

Table D2. GC-FIMS by #C distribution report normalized for >200°C fractions of sample 2

HC Type / #C	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.03**	**1.83**	**8.93**	**12.95**	**16.60**	**16.30**	**12.31**	**3.10**	**1.01**	**0.46**	**73.52**
Paraffins	**0.00**	**0.00**	**3.64**	**5.39**	**7.73**	**7.49**	**6.78**	**1.32**	**0.83**	**0.42**	**33.60**
isoparaffins	0.00	0.00	1.73	3.05	5.00	2.62	4.54	0.57	0.60	0.31	18.43
n-Paraffins	0.00	0.00	1.91	2.35	2.72	4.87	2.24	0.75	0.23	0.11	15.18
Cycloparaffins	**0.03**	**1.83**	**5.29**	**7.56**	**8.88**	**8.81**	**5.53**	**1.77**	**0.17**	**0.04**	**39.92**
Monocycloparaffins	0.00	0.42	2.51	3.75	4.33	4.08	2.81	1.00	0.12	0.04	19.06
Dicycloparaffins	0.00	1.32	2.41	2.84	2.94	3.10	1.64	0.50	0.03	0.00	14.79
Polycycloparaffins	0.03	0.09	0.38	0.96	1.60	1.64	1.08	0.27	0.02	0.00	6.07
Aromatics	**1.22**	**2.51**	**5.13**	**6.34**	**5.34**	**3.47**	**1.71**	**0.65**	**0.11**	**0.00**	**26.48**
MonoAromatics	**1.15**	**2.09**	**3.56**	**4.59**	**4.27**	**2.92**	**1.53**	**0.60**	**0.11**	**0.00**	**20.83**
Alkylbenzenes	0.92	0.91	0.85	1.08	1.20	0.97	0.51	0.22	0.04	0.00	6.70
Benzocycloalkanes	0.23	1.17	2.64	3.12	2.31	1.24	0.56	0.21	0.04	0.00	11.52
Benzodicycloalkanes	0.00	0.01	0.07	0.40	0.77	0.71	0.45	0.17	0.02	0.00	2.60
Diaromatics	**0.07**	**0.38**	**1.57**	**1.75**	**1.05**	**0.53**	**0.18**	**0.05**	**0.00**	**0.00**	**5.58**
Naphthalenes	0.07	0.38	1.40	1.19	0.42	0.15	0.04	0.03	0.00	0.00	3.69
Biphenyls	0.00	0.00	0.15	0.41	0.48	0.27	0.12	0.01	0.00	0.00	1.45
Naphthocycloalkanes	0.00	0.00	0.02	0.10	0.08	0.05	0.00	0.00	0.00	0.00	0.25
Fluorenes	0.00	0.00	0.00	0.04	0.07	0.06	0.02	0.00	0.00	0.00	0.19
Triaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**
Phenanthrenes	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.01
Phenanthrocyclolalkn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic Sulfur	**0.00**	**0.04**	**0.00**	**0.00**	**0.01**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.06**
Benzothiophenes	0.00	0.04	0.00	0.00	0.01	0.01	0.00	0.00	0.00	0.00	0.06
Dibenzothiophenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table D3. PIONA report for sample 3 (<200°C fractions)

#C	Naphthenes	Iso-Paraffins	n-Paraffins	Aromatics	Totals
3	0.000	0.000	0.000	0.000	0.000
4	0.000	0.000	0.000	0.000	0.000
5	0.000	0.000	0.000	0.000	0.000
6	0.065	0.019	0.011	0.008	0.103
7	0.319	0.065	0.034	0.046	0.463
8	0.957	0.182	0.076	0.612	1.827
9	4.456	1.090	0.494	2.279	8.319
10	4.740	4.228	0.703	1.010	10.681
11	2.230	3.339	0.589	0.000	6.157
Totals	12.552	9.366	2.066	3.565	27.550

Table D4. GC-FIMS by #C distribution report normalized for >200°C fractions of sample 3

HC Type / #C	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.03**	**7.77**	**18.29**	**15.09**	**12.79**	**10.18**	**8.09**	**4.98**	**2.00**	**0.27**	**79.51**
Paraffins	**0.00**	**0.00**	**2.84**	**1.63**	**0.81**	**0.13**	**0.12**	**0.03**	**0.02**	**0.00**	**5.58**
isoparaffins	0.00	0.00	2.45	1.41	0.65	0.00	0.00	0.00	0.00	0.00	4.52
n-Paraffins	0.00	0.00	0.39	0.23	0.15	0.13	0.12	0.03	0.02	0.00	1.07
Cycloparaffins	**0.03**	**7.77**	**15.45**	**13.46**	**11.99**	**10.06**	**7.96**	**4.95**	**1.98**	**0.27**	**73.92**
Monocycloparaffins	0.00	0.97	6.02	5.21	4.00	2.78	2.16	1.31	0.47	0.05	22.98
Dicycloparaffins	0.02	6.41	8.30	5.60	4.08	3.38	2.85	1.90	0.88	0.19	33.60
Polycycloparaffins	0.02	0.39	1.13	2.65	3.91	3.90	2.94	1.74	0.62	0.04	17.34
Aromatics	**1.35**	**3.71**	**3.63**	**3.14**	**2.88**	**2.37**	**1.81**	**1.23**	**0.35**	**0.03**	**20.49**
MonoAromatics	**1.33**	**3.63**	**3.50**	**3.04**	**2.59**	**2.16**	**1.75**	**1.01**	**0.35**	**0.03**	**19.28**
Alkylbenzenes	0.23	1.16	1.15	1.01	0.73	0.60	0.58	0.41	0.20	0.02	6.10
Benzocycloalkanes	1.11	2.45	2.25	1.48	1.07	0.89	0.70	0.42	0.14	0.00	10.50
Benzodicycloalkanes	0.00	0.02	0.09	0.46	0.79	0.67	0.47	0.18	0.02	0.00	2.68
Diaromatics	**0.02**	**0.07**	**0.13**	**0.20**	**0.29**	**0.21**	**0.05**	**0.22**	**0.00**	**0.00**	**1.19**
Naphthalenes	0.02	0.07	0.12	0.10	0.07	0.04	0.03	0.22	0.00	0.00	0.66
Biphenyls	0.00	0.00	0.01	0.01	0.06	0.10	0.03	0.00	0.00	0.00	0.19
Naphthocycloalkanes	0.00	0.00	0.00	0.09	0.17	0.06	0.00	0.00	0.00	0.00	0.33
Fluorenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Triaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Phenanthrenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenanthrocyclolalkn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic Sulfur	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.01**	**0.00**	**0.00**	**0.00**	**0.02**
Benzothiophenes	0.00	0.01	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.02
Dibenzothiophenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table D5. PIONA report for sample 4 (<200°C fractions)

#C	Naphthenes	Iso-Paraffins	n-Paraffins	Aromatics	Totals
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00
6	0.02	0.00	0.00	0.00	0.03
7	0.37	0.03	0.07	0.17	0.63
8	0.97	0.40	0.33	0.97	2.67
9	1.59	1.38	0.66	2.69	6.32
10	1.92	2.00	0.87	1.58	6.37
11	1.29	2.29	0.92	0.00	4.51
Totals	6.08	6.46	3.10	4.89	20.52

Table D6. GC-FIMS by #C distribution report normalized for >200°C fractions of sample 4

HC Type / #C	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.09**	**3.92**	**16.74**	**18.62**	**14.54**	**10.39**	**6.17**	**2.76**	**0.81**	**0.13**	**74.17**
Paraffins	**0.00**	**0.00**	**3.44**	**2.55**	**1.53**	**0.64**	**0.25**	**0.04**	**0.03**	**0.00**	**8.48**
isoparaffins	0.00	0.00	2.60	1.95	1.21	0.37	0.09	0.00	0.00	0.00	6.23
n-Paraffins	0.00	0.00	0.84	0.60	0.32	0.27	0.16	0.04	0.03	0.00	2.25
Cycloparaffins	**0.09**	**3.92**	**13.30**	**16.07**	**13.01**	**9.75**	**5.92**	**2.72**	**0.79**	**0.13**	**65.69**
Monocycloparaffins	0.00	0.46	3.86	4.95	3.86	2.90	1.77	0.79	0.15	0.02	18.76
Dicycloparaffins	0.02	2.72	7.42	8.13	6.04	4.15	2.32	1.07	0.35	0.08	32.29
Polycycloparaffins	0.08	0.73	2.03	2.99	3.11	2.70	1.83	0.86	0.29	0.03	14.64
Aromatics	**1.34**	**3.63**	**4.08**	**4.39**	**4.29**	**3.50**	**2.43**	**1.38**	**0.64**	**0.16**	**25.83**
MonoAromatics	**1.32**	**3.54**	**3.92**	**4.13**	**3.97**	**3.20**	**2.24**	**1.29**	**0.64**	**0.16**	**24.41**
Alkylbenzenes	0.60	1.77	1.83	1.71	1.35	1.04	0.68	0.41	0.21	0.07	9.68
Benzocycloalkanes	0.73	1.76	2.02	2.07	1.95	1.42	0.91	0.49	0.25	0.07	11.68
Benzodicycloalkanes	0.00	0.01	0.07	0.34	0.67	0.73	0.65	0.38	0.17	0.02	3.06
Diaromatics	**0.01**	**0.05**	**0.15**	**0.25**	**0.32**	**0.30**	**0.18**	**0.09**	**0.00**	**0.00**	**1.37**
Naphthalenes	0.01	0.05	0.10	0.11	0.09	0.06	0.06	0.06	0.00	0.00	0.55
Biphenyls	0.00	0.00	0.05	0.11	0.17	0.18	0.12	0.03	0.00	0.00	0.66
Naphthocycloalkanes	0.00	0.00	0.00	0.03	0.04	0.03	0.00	0.00	0.00	0.00	0.09
Fluorenes	0.00	0.00	0.00	0.01	0.02	0.03	0.01	0.00	0.00	0.00	0.06
Triaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Phenanthrenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenanthrocyclolalkn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic Sulfur	**0.00**	**0.04**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.05**
Benzothiophenes	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05
Dibenzothiophenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table D7. PIONA report for sample 5 (<200°C fractions)

#C	Naphthenes	Iso-Paraffins	n-Paraffins	Aromatics	Totals
3	0.00	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00	0.00
5	0.00	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00
7	0.00	0.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.14	0.14
9	0.42	0.28	0.13	3.07	3.91
10	0.73	1.63	0.83	5.58	8.76
11	0.39	1.84	0.97	0.00	3.19
Totals	1.56	4.06	2.14	8.24	16.00

Table D8. GC-FIMS by #C distribution normalized for >200°C fractions of sample 5

HC Type / #C	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.08**	**2.52**	**11.42**	**12.09**	**12.43**	**10.71**	**10.38**	**2.82**	**0.56**	**0.00**	**63.01**
Paraffins	**0.00**	**0.03**	**4.88**	**5.75**	**6.33**	**4.91**	**6.66**	**1.60**	**0.34**	**0.00**	**30.50**
isoparaffins	0.00	0.00	3.25	4.43	4.89	1.98	5.26	1.74	0.21	0.00	21.74
n-Paraffins	0.00	0.03	1.63	1.33	1.44	2.93	1.40	0.00	0.00	0.00	8.76
Cycloparaffins	**0.08**	**2.49**	**6.55**	**6.33**	**6.10**	**5.80**	**3.72**	**1.22**	**0.22**	**0.00**	**32.51**
Monocycloparaffins	0.00	0.52	2.73	2.79	2.55	2.32	1.90	0.60	0.00	0.00	13.40
Dicycloparaffins	0.01	1.65	3.03	2.47	2.15	2.15	1.61	0.00	0.00	0.00	13.06
Polycycloparaffins	0.07	0.33	0.78	1.08	1.41	1.34	1.04	0.00	0.00	0.00	6.05
Aromatics	**1.98**	**7.71**	**10.17**	**8.72**	**4.97**	**2.36**	**0.90**	**0.17**	**0.01**	**0.00**	**36.99**
MonoAromatics	**1.79**	**6.55**	**7.56**	**6.02**	**3.92**	**2.18**	**0.83**	**0.17**	**0.01**	**0.00**	**29.02**
Alkylbenzenes	0.53	1.86	1.65	1.12	0.89	0.82	0.37	0.00	0.00	0.00	7.24
Benzocycloalkanes	1.26	4.67	5.72	4.26	2.25	1.04	0.31	0.00	0.00	0.00	19.50
Benzodicycloalkanes	0.00	0.02	0.19	0.65	0.78	0.65	0.01	0.00	0.00	0.00	2.28
Diaromatics	**0.18**	**1.09**	**2.60**	**2.68**	**1.04**	**0.18**	**0.07**	**0.00**	**0.00**	**0.00**	**7.84**
Naphthalenes	0.18	1.09	2.34	1.84	0.56	0.01	0.00	0.00	0.00	0.00	6.03
Biphenyls	0.00	0.00	0.22	0.47	0.53	0.16	0.00	0.00	0.00	0.00	1.38
Naphthocycloalkanes	0.00	0.00	0.03	0.30	0.00	0.00	0.00	0.00	0.00	0.00	0.33
Fluorenes	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.11
Triaromatics	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
Phenanthrenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Phenanthrocyclolalkn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Aromatic Sulfur	**0.01**	**0.07**	**0.01**	**0.01**	**0.02**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.13**
Benzothiophenes	0.01	0.07	0.01	0.01	0.02	0.01	0.00	0.00	0.00	0.00	0.13
Dibenzothiophenes	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table D9. Integrated results from PIONA and GC-FIMS measurements of sample 2

HC Type / #C	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.002**	**0.023**	**0.152**	**0.588**	**1.302**	**1.898**	**3.820**	**7.988**	**11.59**	**14.85**	**14.59**	**11.02**	**2.770**	**0.900**	**0.413**	**71.904**
Paraffins	**0.002**	**0.010**	**0.045**	**0.262**	**0.787**	**1.317**	**1.949**	**3.255**	**4.825**	**6.913**	**6.700**	**6.070**	**1.183**	**0.747**	**0.373**	**34.436**
isoparaffins	0.000	0.003	0.017	0.122	0.447	0.740	1.057	1.546	2.726	4.477	2.346	4.061	0.514	0.539	0.277	18.872
n-Paraffins	0.002	0.007	0.028	0.140	0.340	0.577	0.892	1.709	2.099	2.436	4.354	2.008	0.668	0.208	0.096	15.564
Cycloparaffins	**0.000**	**0.013**	**0.107**	**0.327**	**0.515**	**0.581**	**1.871**	**4.733**	**6.765**	**7.942**	**7.886**	**4.947**	**1.587**	**0.153**	**0.040**	**37.468**
Monocycloparaffins	0.000	0.013	0.026	0.082	0.144	0.159	0.612	2.242	3.360	3.873	3.650	2.512	0.891	0.108	0.040	17.713
Dicycloparaffins	0.000	0.000	0.081	0.238	0.353	0.399	1.179	2.152	2.545	2.633	2.772	1.470	0.452	0.029	0.000	14.304
Polycycloparaffins	0.000	0.000	0.000	0.006	0.018	0.023	0.081	0.338	0.860	1.436	1.464	0.965	0.244	0.016	0.000	5.451
Aromatics	**0.000**	**0.005**	**0.052**	**0.358**	**0.892**	**4.180**	**2.249**	**4.595**	**5.672**	**4.779**	**3.106**	**1.529**	**0.578**	**0.098**	**0.003**	**28.095**
MonoAromatics	**0.000**	**0.005**	**0.052**	**0.358**	**0.892**	**4.119**	**1.873**	**3.183**	**4.107**	**3.825**	**2.616**	**1.368**	**0.534**	**0.098**	**0.003**	**23.034**
Alkylbenzenes	0.000	0.005	0.052	0.358	0.683	2.619	0.818	0.760	0.963	1.070	0.866	0.459	0.197	0.040	0.001	8.892
Benzocycloalkanes	0.000	0.000	0.000	0.000	0.209	1.500	1.046	2.362	2.788	2.070	1.112	0.505	0.186	0.037	0.001	11.815
Benzodicycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.010	0.061	0.356	0.685	0.638	0.404	0.150	0.021	0.001	2.327
Diaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.061**	**0.340**	**1.409**	**1.563**	**0.943**	**0.476**	**0.159**	**0.044**	**0.000**	**0.000**	**4.994**
Naphthalenes	0.000	0.000	0.000	0.000	0.000	0.061	0.340	1.256	1.068	0.376	0.134	0.039	0.030	0.000	0.000	3.303
Biphenyls	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.138	0.371	0.431	0.245	0.103	0.013	0.000	0.000	1.301
Naphthocycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.015	0.091	0.073	0.045	0.001	0.000	0.000	0.000	0.224
Fluorenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.033	0.064	0.053	0.016	0.000	0.000	0.000	0.166
Triaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.006**	**0.006**	**0.000**	**0.000**	**0.000**	**0.000**	**0.012**
Phenanthrenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.006	0.000	0.000	0.000	0.000	0.012
Phenanthrocyclolalkn	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Aromatic Sulfur	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.036**	**0.002**	**0.002**	**0.006**	**0.007**	**0.002**	**0.000**	**0.000**	**0.000**	**0.055**
Benzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.036	0.002	0.002	0.006	0.004	0.001	0.000	0.000	0.000	0.052
Dibenzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.003

Table D10. Integrated results from PIONA and GC-FIMS measurements of sample 3

HC Type / #C	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.000**	**0.095**	**0.418**	**1.215**	**6.040**	**9.694**	**11.79**	**13.25**	**10.94**	**9.269**	**7.378**	**5.859**	**3.607**	**1.447**	**0.199**	**81.198**
Paraffins	**0.000**	**0.030**	**0.099**	**0.258**	**1.584**	**4.930**	**3.928**	**2.058**	**1.184**	**0.585**	**0.093**	**0.090**	**0.022**	**0.014**	**0.000**	**14.876**
isoparaffins	0.000	0.019	0.065	0.182	1.090	4.228	3.339	1.777	1.021	0.474	0.000	0.000	0.000	0.000	0.000	12.195
n-Paraffins	0.000	0.011	0.034	0.076	0.494	0.703	0.589	0.281	0.163	0.111	0.093	0.090	0.022	0.014	0.000	2.680
Cycloparaffins	**0.000**	**0.065**	**0.319**	**0.957**	**4.456**	**4.763**	**7.862**	**11.19**	**9.751**	**8.684**	**7.285**	**5.769**	**3.585**	**1.433**	**0.199**	**66.322**
Monocycloparaffins	0.000	0.065	0.115	0.345	1.646	1.796	2.934	4.360	3.778	2.896	2.015	1.567	0.949	0.343	0.037	22.845
Dicycloparaffins	0.000	0.000	0.204	0.612	2.801	2.956	4.646	6.017	4.054	2.954	2.445	2.068	1.375	0.638	0.135	30.904
Polycycloparaffins	0.000	0.000	0.000	0.001	0.008	0.012	0.282	0.818	1.920	2.835	2.824	2.134	1.261	0.452	0.026	12.573
Aromatics	**0.000**	**0.008**	**0.046**	**0.612**	**2.279**	**1.989**	**2.686**	**2.628**	**2.274**	**2.089**	**1.715**	**1.313**	**0.891**	**0.253**	**0.020**	**18.802**
MonoAromatics	**0.000**	**0.008**	**0.046**	**0.612**	**2.279**	**1.977**	**2.628**	**2.533**	**2.130**	**1.875**	**1.563**	**1.268**	**0.733**	**0.253**	**0.020**	**17.924**
Alkylbenzenes	0.000	0.008	0.046	0.612	1.313	0.554	0.843	0.836	0.730	0.529	0.438	0.421	0.301	0.141	0.018	6.787
Benzocycloalkanes	0.000	0.000	0.000	0.000	0.966	1.423	1.773	1.630	1.070	0.778	0.642	0.509	0.302	0.099	0.002	9.194
Benzodicycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.012	0.067	0.330	0.569	0.483	0.337	0.131	0.013	0.000	1.943
Diaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.012**	**0.052**	**0.094**	**0.145**	**0.213**	**0.150**	**0.038**	**0.156**	**0.000**	**0.000**	**0.861**
Naphthalenes	0.000	0.000	0.000	0.000	0.000	0.012	0.052	0.088	0.073	0.048	0.032	0.019	0.156	0.000	0.000	0.481
Biphenyls	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.005	0.041	0.071	0.019	0.000	0.000	0.000	0.138
Naphthocycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.003	0.067	0.124	0.047	0.000	0.000	0.000	0.000	0.240
Fluorenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001	0.000	0.000	0.000	0.000	0.002
Triaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Phenanthrenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phenanthrocyclolalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Aromatic Sulfur	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.005**	**0.000**	**0.000**	**0.000**	**0.002**	**0.007**	**0.002**	**0.000**	**0.000**	**0.016**
Benzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.000	0.002	0.007	0.002	0.000	0.000	0.016
Dibenzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table D11. Integrated results from PIONA and GC-FIMS measurements of sample 4

HC Type / #C	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.000**	**0.023**	**0.463**	**1.702**	**3.628**	**4.857**	**7.620**	**13.31**	**14.80**	**11.55**	**8.257**	**4.902**	**2.190**	**0.645**	**0.103**	**74.053**
Paraffins	**0.000**	**0.000**	**0.098**	**0.731**	**2.042**	**2.875**	**3.215**	**2.735**	**2.028**	**1.218**	**0.512**	**0.198**	**0.030**	**0.020**	**0.000**	**15.701**
isoparaffins	0.000	0.000	0.032	0.404	1.377	2.001	2.294	2.067	1.553	0.960	0.298	0.072	0.000	0.000	0.000	11.059
n-Paraffins	0.000	0.000	0.066	0.327	0.665	0.874	0.921	0.667	0.475	0.258	0.214	0.126	0.030	0.020	0.000	4.643
Cycloparaffins	**0.000**	**0.023**	**0.365**	**0.971**	**1.586**	**1.982**	**4.405**	**10.57**	**12.77**	**10.34**	**7.745**	**4.703**	**2.160**	**0.625**	**0.103**	**58.352**
Monocycloparaffins	0.000	0.023	0.191	0.449	0.675	0.816	1.663	3.067	3.937	3.065	2.301	1.406	0.630	0.120	0.015	18.357
Dicycloparaffins	0.000	0.000	0.175	0.516	0.885	1.117	2.162	5.893	6.464	4.800	3.297	1.843	0.851	0.275	0.064	28.342
Polycycloparaffins	0.000	0.000	0.000	0.007	0.026	0.049	0.580	1.613	2.373	2.472	2.147	1.454	0.680	0.230	0.024	11.653
Aromatics	**0.000**	**0.005**	**0.166**	**0.967**	**2.691**	**2.640**	**2.885**	**3.239**	**3.486**	**3.411**	**2.781**	**1.930**	**1.094**	**0.509**	**0.130**	**25.936**
MonoAromatics	**0.000**	**0.005**	**0.166**	**0.967**	**2.691**	**2.627**	**2.811**	**3.116**	**3.284**	**3.155**	**2.543**	**1.784**	**1.021**	**0.506**	**0.130**	**24.809**
Alkylbenzenes	0.000	0.005	0.166	0.967	1.962	1.378	1.405	1.453	1.363	1.071	0.830	0.537	0.329	0.170	0.058	11.693
Benzocycloalkanes	0.000	0.000	0.000	0.000	0.729	1.250	1.397	1.607	1.648	1.553	1.132	0.727	0.391	0.198	0.053	10.687
Benzodicycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.009	0.056	0.273	0.531	0.582	0.520	0.302	0.138	0.019	2.429
Diaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.012**	**0.040**	**0.121**	**0.202**	**0.253**	**0.236**	**0.145**	**0.072**	**0.003**	**0.000**	**1.086**
Naphthalenes	0.000	0.000	0.000	0.000	0.000	0.012	0.040	0.082	0.087	0.075	0.051	0.044	0.048	0.003	0.000	0.440
Biphenyls	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.037	0.089	0.137	0.143	0.096	0.023	0.001	0.000	0.526
Naphthocycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.002	0.022	0.029	0.021	0.000	0.000	0.000	0.000	0.074
Fluorenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.013	0.022	0.005	0.002	0.000	0.000	0.046
Triaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**
Phenanthrenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.001
Phenanthrocyclolalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Aromatic Sulfur	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.001**	**0.035**	**0.001**	**0.000**	**0.002**	**0.002**	**0.000**	**0.000**	**0.000**	**0.000**	**0.041**
Benzothiophenes	0.000	0.000	0.000	0.000	0.000	0.001	0.035	0.001	0.000	0.002	0.002	0.000	0.000	0.000	0.000	0.041
Dibenzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table D12. Integrated results from PIONA and GC-FIMS measurements of sample 5

HC Type / #C	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	Sum
Saturates	**0.000**	**0.000**	**0.000**	**0.002**	**0.837**	**3.242**	**5.283**	**9.596**	**10.15**	**10.44**	**8.997**	**9.419**	**1.971**	**0.176**	**0.000**	**60.115**
Paraffins	**0.000**	**0.000**	**0.000**	**0.000**	**0.416**	**2.456**	**2.803**	**4.098**	**4.832**	**5.316**	**4.123**	**5.596**	**1.462**	**0.176**	**0.000**	**31.278**
isoparaffins	0.000	0.000	0.000	0.000	0.283	1.631	1.836	2.731	3.717	4.108	1.659	4.417	1.462	0.176	0.000	22.019
n-Paraffins	0.000	0.000	0.000	0.000	0.133	0.826	0.967	1.368	1.114	1.208	2.464	1.179	0.000	0.000	0.000	9.259
Cycloparaffins	**0.000**	**0.000**	**0.000**	**0.002**	**0.421**	**0.786**	**2.481**	**5.498**	**5.320**	**5.125**	**4.874**	**3.822**	**0.509**	**0.000**	**0.000**	**28.837**
Monocycloparaffins	0.000	0.000	0.000	0.001	0.181	0.259	0.820	2.296	2.342	2.139	1.946	1.597	0.504	0.000	0.000	12.087
Dicycloparaffins	0.000	0.000	0.000	0.001	0.216	0.481	1.383	2.543	2.073	1.802	1.805	1.355	0.003	0.000	0.000	11.663
Polycycloparaffins	0.000	0.000	0.000	0.000	0.024	0.045	0.277	0.659	0.905	1.183	1.123	0.870	0.002	0.000	0.000	5.088
Aromatics	**0.000**	**0.000**	**0.000**	**0.137**	**3.073**	**7.236**	**6.474**	**8.541**	**7.325**	**4.176**	**2.264**	**0.635**	**0.004**	**0.000**	**0.000**	**39.864**
MonoAromatics	**0.000**	**0.000**	**0.000**	**0.137**	**3.073**	**7.079**	**5.500**	**6.348**	**5.058**	**3.291**	**2.107**	**0.576**	**0.000**	**0.000**	**0.000**	**33.169**
Alkylbenzenes	0.000	0.000	0.000	0.137	1.016	2.195	1.562	1.387	0.939	0.750	0.688	0.308	0.000	0.000	0.000	8.982
Benzocycloalkanes	0.000	0.000	0.000	0.000	2.057	4.883	3.920	4.804	3.577	1.889	0.875	0.262	0.000	0.000	0.000	22.267
Benzodicycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.018	0.157	0.543	0.652	0.544	0.006	0.000	0.000	0.000	1.920
Diaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.153**	**0.916**	**2.180**	**2.254**	**0.871**	**0.147**	**0.059**	**0.004**	**0.000**	**0.000**	**6.585**
Naphthalenes	0.000	0.000	0.000	0.000	0.000	0.153	0.916	1.968	1.546	0.474	0.004	0.000	0.000	0.000	0.000	5.061
Biphenyls	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.188	0.394	0.444	0.131	0.000	0.000	0.000	0.000	1.157
Naphthocycloalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.024	0.250	0.001	0.000	0.000	0.000	0.000	0.000	0.275
Fluorenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.092	0.001	0.000	0.000	0.000	0.000	0.000	0.093
Triaromatics	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**
Phenanthrenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Phenanthrocyclolalkanes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Aromatic Sulfur	**0.000**	**0.000**	**0.000**	**0.000**	**0.000**	**0.005**	**0.059**	**0.012**	**0.012**	**0.013**	**0.010**	**0.000**	**0.000**	**0.000**	**0.000**	**0.111**
Benzothiophenes	0.000	0.000	0.000	0.000	0.000	0.005	0.059	0.012	0.012	0.013	0.010	0.000	0.000	0.000	0.000	0.111
Dibenzothiophenes	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Appendix E. A comprehensive comparison between this work and reported characterization methods

Characterization method	Number of molecules	Experiments based	Bulk property check	Distribution property check
Monte Carlo Simulation (Neurock et al., 1994)	>10,000 for VGO, HGO and Asphaltene	NMR, VPO, CHNS Reported GC/MS	MW ± 90 amu H/C	SimDis ±25°C SG by Bp (no exp) Arom. Dist. Not consistent
Pseudo-Component (Liguras & Allen, 1989)	325 max. for VGO	GC/MS, NMR	Derived C center distribution	Predict the cracking behavior (no exp)
Structural Orientated Lump (Mobil)	~3000 for VGO (isomeric lumps)	GC/FIMS etc.	FCC Product yield (Gasoline, LCO, Coke)	Predict FCC products (with pilot plant data)
Quadrature Method (Campbell & Klein, 1997)	10-20 for crudes with p.d.f from MC simulation	VPO, SARA, CHNS, NMR	MW, H/C, SARA, S%	4 cuts D86 for Naphtha PINA by #C for Naphtha Not continuous
Structural Group Assembly (Khorasheh et al., 1998)	600 for crude oil	CHNS, NMR (Structural Group Analysis)	Structural Group Distribution from SGA	N/A
Structural Attribute Assembly (Sheremata et al., 2004)	100→6 for Athabasca Asphaltene	CHNSOV, ^{1}H & ^{13}C NMR, VPO	MW, CHNSOV, and Carbon Type from NMR	N/A
Deterministic method **(Thermodynamics applied)**	1000-2000 for Diesels (PM3 simulated)	PIONA, SimDis, GC/FIMS	PIONA, SARA, CHNS, 16 HC Type, MW, Density, RI	SimDis, and 16 HC Type Consistent Distribution (GC-FIMS info conserved)